REMARKS

I. Comments on the Restriction Requirement

Applicants note that Claims 10-11 (Group V), Claim 12 (Group VI), Claim 16 (Group X), and Claim 17 (Group XI) are "method of use" claims drawn to methods of using the polynucleotides of Group II, which should be examined together with the polynucleotide claims of Group II, per the Commissioner's Notice in the Official Gazette of March 26, 1996, entitled "Guidance on Treatment of Product and Process Claims in light of *In re Ochiai, In re Brouwer*, and 35 U.S.C. § 103(b)" which sets forth the rules, upon allowance of product claims, for rejoinder of process claims covering the same scope of products.

II. Priority Information

The Examiner requested that the priority information in the Specification be amended to reflect that U.S. application serial number 09/309,320 had issued as a patent. (Office Action, page 4.) The Specification has been amended accordingly.

III. Publications Cited in the Office Action

The Examiner cited Russell [J. Mol. Bio. 244:332-350], Skolnick et al. [Trends in Biotech. 18(1):34-39], and Attwood [Science 290: 471-473, 29 October 2000] in support of the enablement rejection under 35 U.S.C. § 112, first paragraph (Office Action, page 10.) Applicants note that copies of the Russell, Skolnick et al., and Attwood publications were neither listed on the PTO-892 form nor included with the Office Action.

IV. Objections to Claims 2-4

The Examiner objected to Claims 2-4 because "Claim 2 is dependent on non-elected claim 1." (Office Action, page 5.) Amended Claim 2 is an independent claim. Claims 3 and 4 depend from Claim 2. Therefore, Applicants respectfully request that the Examiner withdraw [illegible]

V. Rejection of Claims 2-4 and 8-9 Under 35 U.S.C. § 112, first paragraph, written description

Claims 2-4 and 8-9 have been rejected under the first paragraph of 35 U.S.C. 112 for alleged lack of an adequate written description. The Examiner alleges that the polynucleotides of Claim 2 encoding polypeptide fragments, the polynucleotide variants of Claim 8, and the polynucleotides of Claim 8 complementary to or ribonucleotide equivalents of SEQ ID NO:2 and SEQ ID NO:2 variants are not adequately described.[1]

Solely in order to expedite prosecution, Applicants have amended Claim 2 such that polynucleotides encoding biologically active or immunogenic fragments of SEQ ID NO:1 are no longer recited. Therefore the rejection as it pertains to polynucleotides encoding biologically active or immunogenic fragments of SEQ ID NO:1 is moot.

The Examiner ignores the claim limitations of "at least 90% identical to a polynucleotide sequence of SEQ ID NO:2" and attempts to introduce a limitation of "function" to the polynucleotide variants, limitations which are not present in the pending claims. The Examiner ignores the limitation that the claimed polynucleotides comprise a naturally occurring polynucleotide sequence.

The requirements necessary to fulfill the written description requirement of 35 U.S.C. 112, first paragraph, are well established by case law.

> . . . the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession *of the invention.* The invention is, for purposes of the "written description" inquiry, *whatever is now claimed. Vas-Cath, Inc. v. Mahurkar,* 19 USPQ2d 1111, 1117 (Fed. Cir. 1991)

Attention is also drawn to the Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001, which provide that :

> An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics[42] which provide evidence that applicant was

> in possession of the claimed invention,[23] i.e., complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics.[24] What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail.[25] If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met.[26]

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art.

SEQ ID NO:1 and SEQ ID NO:2 are specifically disclosed in the application (see, for example, pages 50-52). Variants of SEQ ID NO:1 are described, for example, at page 6, lines 5-14. In particular, the preferred, more preferred, and most preferred SEQ ID NO:1 variants (80%, 90%, and 95% amino acid sequence similarity to SEQ ID NO:1) are described, for example, at page 12, lines 23-26. Variants of SEQ ID NO:2 are described, for example, at page 12, line 27 through page 13, line 19. Incyte clones in which the nucleic acids encoding the human HGST were first identified and libraries from which those clones were isolated are described, for example, at page 11, line 28 through page 12, line 3 of the Specification. Chemical and structural features of HGST are described, for example, on page 12, lines 4-20.

Given SEQ ID NO:1, one of ordinary skill in the art would recognize a polynucleotide encoding a naturally-occurring variant of SEQ ID NO:1 having at least 90% sequence identity to SEQ ID NO:1. Given SEQ ID NO:2 one of ordinary skill in the art would recognize a naturally-occurring variant of SEQ ID NO:2 having at least 90% sequence identity to SEQ ID NO:2. The Specification describes how to use BLAST to determine whether a given sequence falls within the "at least 90% identical" scope. (Specification, page 41, line 28 through page 42, line 13.)

There simply is no requirement that the claims recite particular variant polypeptide, variant polynucleotide, complementary polynucleotide, or ribonucleotide equivalent polynucleotide sequences because the claims already provide sufficient structural definition of the claimed subject matter. That is, [illegible]

occurring amino acid sequence having at least 90% sequence identity to the sequence of SEQ ID NO:1 over the entire length of SEQ ID NO:1." The polynucleotide variants, complementary polynucleotides, and ribonucleotide equivalent polynucleotides are defined in terms of SEQ ID NO:2 ("An isolated polynucleotide comprising a sequence selected from the group consisting of: a) a polynucleotide sequence of SEQ ID NO:2, b) a naturally-occurring polynucleotide sequence having at least 90% sequence identity to the sequence of SEQ ID NO:2, over the entire length of SEQ ID NO:2, c) a polynucleotide sequence completely complementary to a), d) a polynucleotide sequence completely complementary to b) and e) a ribonucleotide equivalent of a)-d).")

Because the recited polypeptide variants are defined in terms of SEQ ID NO:1, and the recited polynucleotide variants, complementary polynucleotides, and ribonucleotide equivalent polynucleotides are defined in terms of SEQ ID NO:2, the precise chemical structure of every polypeptide variant, every polynucleotide variant, every complementary polynucleotide, and every ribonucleotide equivalent polynucleotide within the scope of the claims can be discerned. The Examiner's position is nothing more than a misguided attempt to require Applicants to unduly limit the scope of their claimed invention. Accordingly, the Specification provides an adequate written description of the recited polypeptide and polynucleotide sequences.

A. The present claims specifically define the claimed genus through the recitation of chemical structure

Court cases in which "DNA claims" have been at issue commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed Cir. 1993), the court stated that:

> If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found [illegible]

any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

> In claims to genetic material, however, a generic statement such as "vertebrate insulin cDNA" or "mammalian insulin cDNA," without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

> 1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count:

> A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35 U.S.C. §112; *i.e.*, "an mRNA of a vertebrate, which mRNA encodes insulin" in *Lilly*, and "DNA which codes for a human fibroblast interferon-beta polypeptide" in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polynucleotides in terms of chemical structure, rather than on functional characteristics. For example, the [illegible]

> 2. An isolated polynucleotide encoding a polypeptide comprising an amino acid sequence selected from the group consisting of. . .
>
> b) a naturally-occurring amino acid sequence having at least 90% sequence identity to the sequence of SEQ ID NO:1 over the entire length of SEQ ID NO:1.
>
> 8. An isolated polynucleotide comprising a sequence selected from the group consisting of. . .
>
> b) a naturally-occurring polynucleotide sequence having at least 90% sequence identity to the sequence of SEQ ID NO:2, over the entire length of SEQ ID NO:2. . .

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1 and SEQ ID NO:2. In the present case, there is no reliance merely on a description of functional characteristics of the polynucleotides recited by the claims. In fact, there is no recitation of functional characteristics. Moreover, if such functional recitations were included, it would add to the structural characterization of the recited polynucleotides. The polynucleotides defined in the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base its written description inquiry "on whatever is now claimed," the Office Action failed to provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*.

B. The present claims do not define a genus which is highly diverse

Furthermore, the claims at issue do not describe a genus which could be characterized as highly diverse. Available evidence illustrates that the claimed genus is of narrow scope.

In support of this assertion, the Examiner's attention is directed to the enclosed reference by Brenner et al. ("Assessing sequence comparison methods with reliable structurally identified distant

and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues. (Brenner et al., pages 6073 and 6076.) Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that ~40% identity over at least 70 residues is reliable in signifying homology between proteins. (Brenner et al., page 6076.)

The present application is directed, *inter alia*, to glutathione s-transferase proteins related to the amino acid sequence of SEQ ID NO:1. In accordance with Brenner et al, naturally occurring molecules may exist which could be characterized as glutathione s-transferase proteins and which have as little as 40% identity over at least 70 residues to SEQ ID NO:1. The "variant language" of the present claims recites, for example, polynucleotides encoding a polypeptide comprising "a naturally-occurring amino acid sequence having at least 90% sequence identity to the sequence of SEQ ID NO:1 over the entire length of SEQ ID NO 1." This variation is far less than that of all potential glutathione s-transferase proteins related to SEQ ID NO:1, i.e., those glutathione s-transferase proteins having as little as 40% identity over at least 70 residues to SEQ ID NO:1.

C. The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. §112. The '525 patent claimed the benefit of priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those case was based on the state of the art at essentially at the "dark ages" of recombinant DNA technology.

The present application has a priority date of November 26, 1996. Much has happened in the development of recombinant DNA technology in the 17 or more years from the time of filing of the [illegible]

technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances one of skill in the art would recognize that, given the sequence information of SEQ ID NO:1 and SEQ ID NO:2, and the additional extensive detail provided by the subject application, the present inventors were in possession of the recited polypeptide variants, polynucleotide variants, complementary polynucleotides, and ribonucleotide equivalent polynucleotides at the time of filing of this application.

D. Summary

The Office Action failed to base its written description inquiry "on whatever is now claimed." Consequently, the Action did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1 or SEQ ID NO:2. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polynucleotides defined by the present claims is adequately described, as evidenced by Brenner et al. Furthermore, there have been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Office Action.

VI. Rejection of Claims 2-4 and 8-9 Under 35 U.S.C. § 112, first paragraph, enablement

Applicants' invention is directed, *inter alia*, to polynucleotides encoding polypeptides (HGST) having strong homology to two human Alpha GSTs, pGTH2 (GI 825605; SEQ ID NO:3), and A1-1 (GI 259141; SEQ ID NO:4), and a mouse Alpha GSH, GST 5.7 (GI 193710; SEQ ID NO:5). These polynucleotides have a variety of utilities, in particular in expression profiling, and in particular for

drug discovery (see the Specification at, e.g., page 34, line 18 through page 37, line 11). As described in the Specification (page 11, line 28 through page 12, line 22):

> Nucleic acids encoding the human HGST of the present invention were first identified in Incyte Clone 1553079 from the bladder tumor cDNA library (BLADTUT04) through a computer-generated search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:2, was derived from the following overlapping and/or extended nucleic acid sequences: Incyte Clones 1553079/ BLADTUT04, 1328546/ PANCNOT07, 1422059/ KIDNNOT09, and 2188683/ PROSNOT26.
>
> In one embodiment, the invention encompasses the novel human glutathione s-transferase, a polypeptide comprising the amino acid sequence of SEQ ID NO:1, as shown in Figures 1A, 1B, and 1C. HGST is 222 amino acids in length and has chemical and structural homology with two human Alpha GSTs, pGTH2 (GI 825605; SEQ ID NO:3), and A1-1 (GI 259141), and a mouse Alpha GSH, GST 5.7 (GI 193710). In particular, HGST shares 57% overall identity with each of the two human GSTs and 59% identity with the mouse GST. In addition, various amino acid residues found to be essential for the catalytic activity and substrate binding of GSTs are conserved in HGST and in the other three GST molecules. These residues are: Y9, R13, R20, E32, Q67, T68, R69, E97, D101, E104, and R131. Only residues E97 and E104 are not found in the mouse GST. E32 and E97 form salt bridges with R20 and R69, respectively, and these salt bridges or residues are thought to be important in structural stability of the GST molecule and may be important for catalysis. Y9 is essential for catalysis by facilitating ionization of GSH. Residues Q67, T68, D101, E104, and R131 are important for the binding of GSH. As illustrated by Figures 3, 4, 5, and 6, HGST and the three Alpha GSTs have rather similar hydrophobicity plots. Figures 7, 8, and 9 show the isoelectric point analyses for HGST, pGHT2, and A1-1. The pI values of 8.8, 9.0, and 9.3, respectively; fall within the range characteristic of Alpha GSTs. In addition to bladder tumor, partial transcripts of the cDNA encoding HGST are found in fetal tissues (kidney and pancreas) and in prostate tissue adjacent to prostate cancer.

Claims 2-4 and 8-9 stand rejected under 35 U.S.C. § 112, first paragraph, based on the allegation that the claimed invention is not adequately enabled. The rejection alleges in particular, regarding the claimed polynucleotide variants of SEQ ID NO:2[2], "absent factual evidence, a percentage

[2] Applicants note that the Examiner did not include the polynucleotides encoding variants and

sequence similarity of less than 100% is not deemed to reasonably support, to one skilled in the art, as to whether the biochemical activity of the claimed subject matter would be the same as that of such a similar known biomolecule." (Office Action, page 9.) The Examiner further alleged that the claims contain "subject matter which was not described in the specification in such a way as to enable one of skill in the art to which it pertains, or with which it is most nearly connected, to make and/or use the invention" and that "a skilled artisan would be forced into undue experimentation to practice (i.e., make and use) the invention as is broadly claimed." (Office Action, pages 9-10.)

The Examiner further alleged that "[t]he need for non-routine experimentation demonstrates that the specification is not enabled for any asserted use or well-established use for bacterial membrane polypeptides."[3] (Office Action, page 10.)

The claimed polynucleotides are enabled, i.e., they are supported by the Specification and what is well known in the art.

A. How to make

SEQ ID NO:1 and SEQ ID NO:2 are specifically disclosed in the application (see, for example, pages 50-52 of the Sequence Listing). Variants of SEQ ID NO:1 are disclosed, for example, at page 6, lines 6-14. In particular, the preferred, more preferred, and most preferred SEQ ID NO:1 variants (80%, 90%, and 95% amino acid sequence similarity to SEQ ID NO:1) are disclosed, for example, at page 12, lines 23-26. Variants of SEQ ID NO:2 are disclosed, for example, at page 12, line 27 through page 13, line 19 and page 13, line 26 through page 14, line 21. Incyte clones in which the nucleic acids encoding the human HGST were first identified and libraries from which those clones were isolated are disclosed, for example, at page 11, line 18 through page 12, line 3 of the Specification. Chemical and structural features of HGST are disclosed, for example, on page 12, lines 4-20.

[3] Applicants note that the specification and claims do not recite "bacterial membrane

The Examiner alleged that the claims were not enabled because "[i]n absence of further guidance from Applicants, the skilled artisan would have to de novo discover what the appropriate conservative amino acid substitutions are and what critical regions can and cannot tolerate such substitutions" and "absent factual evidence, a percentage sequence similarity of less than 100% is not deemed to reasonably support, to one skilled in the art, as to whether the biological activity of the claimed subject matter would be the same as that of such a similar known biomolecule." (Office Action, pages 9 and 10.) However, Applicants submit that the polypeptide variant sequences and polynucleotide variant sequences are defined by their being "naturally occurring" and by their percentage sequence identity with SEQ ID NO:1 and SEQ ID NO:2 and not by biological function. The choice of amino acids or nucleotides to alter is made by nature. "Naturally occurring" polypeptide variant sequences and polynucleotide variant sequences occur in nature; they are not created exclusively in a laboratory. The Specification teaches how to find polynucleotide variants (e.g., page 35, lines 3-15) which can then be expressed to make polypeptide variants and how to use BLAST methods to determine whether a given naturally occurring polynucleotide sequence falls within the "at least 90% identical to a polynucleotide sequence of SEQ ID NO:2" scope and whether a given naturally occurring amino acid sequence falls within the "at least 90% identical to an amino acid sequence of SEQ ID NO:1" scope (e.g., page 41, line 28 through page 42, line 13). In addition, determination of percent identity is well known in the art.

The making of the claimed polynucleotides by recombinant and chemical synthetic methods is disclosed in the Specification, at, e.g., page 13, lines 20-25, page 16, lines 16-21, and page 17, lines 13-15.

This satisfies the "how to make" requirement of 35 U.S.C. § 112, first paragraph.

B. How to Use

The rejection of Claims 2-4 and 8-9 is improper, as the inventions of those claims are [illegible]

The invention at issue is a polynucleotide sequence corresponding to a gene that is expressed in human bladder tumor tissue, as well as polynucleotides encoding SEQ ID NO:1 variants, variants of the SEQ ID NO:2 polynucleotide, complementary polynucleotides, and ribonucleotide equivalents polynucleotides (hereinafter "the claimed polynucleotides"). The SEQ ID NO:2 polynucleotide codes for a polypeptide demonstrated in the patent specification to be a member of the class of glutathione s-transferases, whose biological functions include deactivation and detoxification of potentially mutagenic and carcinogenic chemicals. (Specification, pages 1-3.) The claimed invention has numerous practical, beneficial uses in toxicology testing, drug development, and the diagnosis of disease, none of which requires knowledge of how the polypeptides coded for by the claimed polynucleotides actually function. As a result of the benefits of these uses, the claimed invention already enjoys significant commercial success.

Applicants submit with this Response the Declaration of Dr. Tod Bedilion[4] describing some of the practical uses of the claimed invention in gene and protein expression monitoring applications. The Bedilion Declaration demonstrates that the positions and arguments made by the Patent Examiner with respect to the enablement and utility of the claimed polynucleotides are without merit.

The Bedilion Declaration describes, in particular, how the claimed expressed polynucleotides can be used in gene expression monitoring applications that were well-known at the time the patent application was filed, and how those applications are useful in developing drugs and monitoring their activity. Dr. Bedilion states that the claimed invention is a useful tool when employed as a highly specific probe in a cDNA microarray:

> Persons skilled in the art would [have appreciated on November 26, 1996] that cDNA microarrays that contained the claimed polynucleotides would be a more useful tool than cDNA microarrays that did not contain the polynucleotides in connection with conducting gene expression monitoring studies on proposed (or actual) drugs for treating cancer for such purposes as evaluating their efficacy and toxicity. (Bedilion Declaration ¶ 15.)

The Patent Examiner contends that the claimed polynucleotides cannot be useful without precise knowledge of their biological function. But the law never has required knowledge of biological function to prove utility. It is the claimed invention's uses, not its functions, that are the subject of a proper analysis under the utility requirement.

In any event, as demonstrated by the Bedilion Declaration, the person of ordinary skill in the art can achieve beneficial results from the claimed polynucleotides in the absence of any knowledge as to the precise function of the proteins encoded by them. The uses of the claimed polynucleotides in gene expression monitoring applications are in fact independent of their precise function.

1. The Applicable Legal Standard

To meet the utility requirement of sections 101 and 112 of the Patent Act, the patent applicant need only show that the claimed invention is "practically useful," *Anderson v. Natta*, 480 F.2d 1392, 1397, 178 USPQ 458 (CCPA 1973) and confers a "specific benefit" on the public. *Brenner v. Manson*, 383 U.S. 519, 534-35, 148 USPQ 689 (1966). As discussed in a recent Court of Appeals for the Federal Circuit case, this threshold is not high:

> An invention is "useful" under section 101 if it is capable of providing some identifiable benefit. See *Brenner v. Manson*, 383 U.S. 519, 534 [148 USPQ 689] (1966); *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 [24 USPQ2d 1401] (Fed. Cir. 1992) ("to violate Section 101 the claimed device must be totally incapable of achieving a useful result"); *Fuller v. Berger*, 120 F. 274, 275 (7th Cir. 1903) (test for utility is whether invention "is incapable of serving any beneficial end").

Juicy Whip Inc. v. Orange Bang Inc., 51 USPQ2d 1700 (Fed. Cir. 1999).

While an asserted utility must be described with specificity, the patent applicant need not demonstrate utility to a certainty. In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180, 20 USPQ2d 1094 (Fed. Cir. 1991), the United States Court of Appeals for the Federal Circuit explained:

> An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not

The specificity requirement is not, therefore, an onerous one. If the asserted utility is described so that a person of ordinary skill in the art would understand how to use the claimed invention, it is sufficiently specific. *See Standard Oil Co. v. Montedison, S.p.a.*, 212 U.S.P.Q. 327, 343 (3d Cir. 1981). The specificity requirement is met unless the asserted utility amounts to a "nebulous expression" such as "biological activity" or "biological properties" that does not convey meaningful information about the utility of what is being claimed. *Cross v. Iizuka*, 753 F.2d 1040, 1048 (Fed. Cir. 1985).

In addition to conferring a specific benefit on the public, the benefit must also be "substantial." *Brenner*, 383 U.S. at 534. A "substantial" utility is a practical, "real-world" utility. *Nelson v. Bowler*, 626 F.2d 853, 856, 206 USPQ 881 (CCPA 1980).

If persons of ordinary skill in the art would understand that there is a "well-established" utility for the claimed invention, the threshold is met automatically and the applicant need not make any showing to demonstrate utility. Manual of Patent Examination Procedure at § 706.03(a). Only if there is no "well-established" utility for the claimed invention must the applicant demonstrate the practical benefits of the invention. *Id.*

Once the patent applicant identifies a specific utility, the claimed invention is presumed to possess it. *In re Cortright*, 165 F.3d 1353, 1357, 49 USPQ2d 1464 (Fed. Cir. 1999); *In re Brana*, 51 F.3d 1560, 1566; 34 USPQ2d 1436 (Fed. Cir. 1995). In that case, the Patent Office bears the burden of demonstrating that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved by the claimed invention. *Id.* To do so, the Patent Office must provide evidence or sound scientific reasoning. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566. The applicant need only prove a "substantial likelihood" of utility; certainty is not required. *Brenner*, 383 U.S. at 532.

2. Uses of the claimed polynucleotides for diagnosis of conditions and disorders characterized by expression of HGST, for toxicology testing,

The claimed invention meets all of the necessary requirements for establishing a credible utility under the Patent Law: There are "well-established" uses for the claimed invention known to persons of ordinary skill in the art, and there are specific practical and beneficial uses for the invention disclosed in the patent application's specification. These uses are explained, in detail, in the Bedilion Declaration accompanying this Response. Objective evidence, not considered by the Patent Office, further corroborates the credibility of the asserted utilities.

a. The uses of the claimed polynucleotides for toxicology testing, drug discovery, and disease diagnosis are practical uses that confer "specific benefits" to the public

The claimed invention has specific, substantial, real-world utility by virtue of its use in toxicology testing, drug development and disease diagnosis through gene expression profiling. These uses are explained in detail in the accompanying Bedilion Declaration. The claimed invention is a useful tool in cDNA microarrays used to perform gene expression analysis. That is sufficient to establish utility for the claimed polynucleotides.

In his Declaration, Dr. Bedilion explains the many reasons why a person skilled in the art reading the Goli '771 application on November 26, 1996 would have understood that application to disclose the claimed polynucleotides to be useful for a number of gene expression monitoring applications, *e.g.*, as highly specific probes for the expression of those specific polynucleotides in connection with the development of drugs and the monitoring of the activity of such drugs. (Bedilion Declaration at, e.g., ¶¶ 10-15). Much, but not all, of Dr. Bedilion's explanation concerns the use of the claimed polynucleotides in cDNA microarrays of the type first developed at Stanford University for evaluating the efficacy and toxicity of drugs, as well as for other applications. (Bedilion Declaration, ¶¶ 12 and 15).[5]

[5] Dr. Bedilion also explained, for example, why persons skilled in the art would also appreciate, [illegible] the Goli '771 application that the claimed polynucleotides would be useful in connection

In connection with his explanations, Dr. Bedilion states that the "Goli '771 application would have led a person skilled in the art on November 26, 1996 who was using gene expression monitoring in connection with working on developing new drugs for the treatment of cancer to conclude that a cDNA microarray that contained the claimed polynucleotides would be a highly useful tool and to request specifically that any cDNA microarray that was being used for such purposes contain the claimed polynucleotides." (Bedilion Declaration, ¶ 15). For example, as explained by Dr. Bedilion, "[p]ersons skilled in the art would [have appreciated on November 26, 1996] that cDNA microarrays that contained the claimed polynucleotides would be a more useful tool than cDNA microarrays that did not contain the polynucleotides in connection with conducting gene expression monitoring studies on proposed (or actual) drugs for treating cancer for such purposes as evaluating their efficacy and toxicity." *Id.*

In support of those statements, Dr. Bedilion provided detailed explanations of how cDNA technology can be used to conduct gene expression monitoring evaluations, with extensive citations to pre-November 26, 1996 publications showing the state of the art on November 26, 1996. (Bedilion Declaration, ¶ ¶ 10-14). While Dr. Bedilion's explanations in paragraph 15 of his Declaration include three pages of text and six subparts (a)-(f), he specifically states that his explanations are not "all-inclusive." *Id.* For example, with respect to toxicity evaluations, Dr. Bedilion had earlier explained how persons skilled in the art who were working on drug development on November 26, 1996 (and for several years prior to November 26, 1996) "without any doubt" appreciated that the toxicity (or lack of toxicity) of any proposed drug was "one of the most important criteria to be evaluated in connection with the development of the drug" and how the teachings of the Goli '771 application clearly include using differential gene expression analyses in toxicity studies (Bedilion Declaration, ¶ 10).

Thus, the Bedilion Declaration establishes that persons skilled in the art reading the Goli '771 application at the time it was filed "would have wanted their cDNA microarray to have a [claimed polynucleotide probe] because a microarray that contained such a probe (as compared to one that did not) would provide more useful results in the kind of gene expression monitoring studies using cDNA [illegible]

conclusion that the Goli '771 application disclosed to persons skilled in the art at the time of its filing substantial, specific and credible real-world utilities for the claimed polynucleotides.

Nowhere does the Patent Examiner address the fact that, as described on pages 22 and 35 of the Goli '739 application, the claimed polynucleotides can be used as highly specific probes in, for example, chip based technologies [cDNA microarrays] – probes that without question can be used to measure both the existence and amount of complementary RNA sequences known to be the expression products of the claimed polynucleotides. The claimed invention is not, in that regard, some random sequence whose value as a probe is speculative or would require further research to determine.

Given the fact that the claimed polynucleotides are known to be expressed, their utility as a measuring and analyzing instrument for expression levels is as indisputable as a scale's utility for measuring weight. This use as a measuring tool, regardless of how the expression level data ultimately would be used by a person of ordinary skill in the art, by itself demonstrates that the claimed invention provides an identifiable, real-world benefit that meets the utility requirement. *Raytheon* v. *Roper*, 724 F.2d 951, (Fed. Cir. 1983) (claimed invention need only meet one of its stated objectives to be useful); *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999) (how the invention works is irrelevant to utility); MPEP § 2107 ("Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific, and unquestionable utility (e.g., <u>they are useful in analyzing compounds</u>)" (emphasis added)).

The Bedilion Declaration shows that a number of pre-November 26, 1996 publications confirm and further establish the utility of cDNA microarrays in a wide range of drug development gene expression monitoring applications at the time the Goli '771 application was filed (Bedilion Declaration ¶¶ 10-14; Bedilion Exhibits A-G). Indeed, Brown and Shalon U.S. Patent No. 5,807,522 (the Brown '522 patent, Bedilion Exhibit D), which issued from a patent application filed in June 1995 and was effectively published on December 29, 1995 as a result of the publication of a PCT counterpart application, shows that the Patent Office recognizes the patentable utility of the cDNA technology developed in the early to mid 1990s. As explained by Dr. Bedilion, among other things (Bedilion

> The Brown '522 patent further teaches that the "[m]icroarrays of immobilized nucleic acid sequences prepared in accordance with the invention" can be used in "numerous" genetic applications, including "monitoring of gene expression" applications (see Bedilion Tab D at col. 14, lines 36-42). The Brown '522 patent teaches (a) monitoring gene expression (i) in different tissue types, (ii) in different disease states, and (iii) in response to different drugs, and (b) that arrays disclosed therein may be used in toxicology studies (see Bedilion Tab D at col. 15, lines 13-18 and 52-58 and col. 18, lines 25-30).

Literature reviews published after the filing of the Goli '771 application describing the state of the art further confirm the claimed invention's utility. Rockett et al. confirm, for example, that the claimed invention is useful for differential expression analysis regardless of how expression is regulated:

> Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years.
>
> * * *
>
> Although differential expression technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up– or down-regulated is required.
>
> * * *
>
> Whereas it would be informative to know the identity and functionality of all genes up/down regulated by . . . toxicants, this would appear a longer term goal . . . . However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. (emphasis in original.)

John C. Rockett, et. al., Differential gene expression in drug metabolism and toxicology: practicalities, problems, and potential, Xenobiotica 29:655-691 (July 1999) (Reference No. 2):

In another post-November 26, 1996 article, Lashkari et al. state explicitly that sequences that are merely "predicted" to be expressed (predicted Open Reading Frames, or ORFs) – the claimed invention in fact is known to be expressed – have numerous uses:

> Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons– they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay.

Lashkari, et al., Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR, (August 1997) Proc. Nat. Acad. Sci. U.S.A. 94:8945-8947 (Reference No. 3). (emphasis added.)

b. The use of nucleic acids coding for proteins expressed by humans as tools for toxicology testing, drug discovery, and the diagnosis of disease is now "well-established"

The technologies made possible by expression profiling and the DNA tools upon which they rely are now well-established. The technical literature recognizes not only the prevalence of these technologies, but also their unprecedented advantages in drug development, testing and safety assessment. These technologies include toxicology testing, as described by Bedilion in his Declaration.

Toxicology testing is now standard practice in the pharmaceutical industry. See, *e.g.*, John C. Rockett et al., (Reference No. 2, *supra*):

> Knowledge of toxin-dependent regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. (Reference No. 2, page 656)

To the same effect are several other scientific publications, including Emile F. Nuwaysir, et al., [illegible] The Advent of Toxicogenomics, Molecular Carcinogenesis 24:153-159

(1999) (Reference No. 4); Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000) (Reference No. 5).

Nucleic acids useful for measuring the expression of whole classes of genes are routinely incorporated for use in toxicology testing. Nuwaysir et al. describes, for example, a Human ToxChip comprising 2089 human clones, which were selected

> for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. (Reference No. 4, page 156.)

See also Table 1 of Nuwaysir et al. (listing additional classes of genes deemed to be of special interest in making a human toxicology microarray).

The more genes that are available for use in toxicology testing, the more powerful the technique. "Arrays are at their most powerful when they contain the entire genome of the species they are being used to study." John C. Rockett and David J. Dix, Application of DNA Arrays to Toxicology, Environ. Health Perspec. 107:681-685 (1999) (Reference No. 6, see page 683). Control genes are carefully selected for their stability across a large set of array experiments in order to best study the effect of toxicological compounds. See attached email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding (Reference No. 7), indicating that even the expression of carefully selected control genes can be altered. Thus, there is no expressed gene which is irrelevant to screening for toxicological effects, and all expressed genes have a utility for toxicological screening.

In fact, the potential benefit to the public, in terms of lives saved and reduced health care costs, are enormous. Recent developments provide evidence that the benefits of this information are already

- In 1999, CV Therapeutics, an Incyte collaborator, was able to use Incyte gene expression technology, information about the structure of a known transporter gene, and chromosomal mapping location, to identify the key gene associated with Tangier disease. This discovery took place over a matter of only a few weeks, due to the power of these new genomics technologies. The discovery received an award from the American Heart Association as one of the top 10 discoveries associated with heart disease research in 1999.

- In an April 9, 2000, article published by the Bloomberg news service, an Incyte customer stated that it had reduced the time associated with target discovery and validation from 36 months to 18 months, through use of Incyte's genomic information database. Other Incyte customers have privately reported similar experiences. The implications of this significant saving of time and expense for the number of drugs that may be developed and their cost are obvious.

- In a February 10, 2000, article in the *Wall Street Journal*, one Incyte customer stated that over 50 percent of the drug targets in its current pipeline were derived from the Incyte database. Other Incyte customers have privately reported similar experiences. By doubling the number of targets available to pharmaceutical researchers, Incyte genomic information has demonstrably accelerated the development of new drugs.

Because the Patent Examiner failed to address or consider the "well-established" utilities for the claimed invention in toxicology testing, drug development, and the diagnosis of disease, the Examiner's rejections should be withdrawn regardless of their merit.

c. The similarity of the polypeptides encoded by the claimed invention to another polypeptide of undisputed utility demonstrates utility

In addition to having substantial, specific and credible utilities in numerous gene expression monitoring applications, the utility of the claimed polynucleotide variants can be imputed based on the relationship between the polypeptides they encode and another polypeptide of unquestioned utility, the SEQ ID NO:1 polypeptide. The polypeptides have sufficient similarities in their sequences that a person of ordinary skill in the art would recognize more than a reasonable probability that the [illegible] similar to the SEQ ID NO:1 polypeptide

It is undisputed, and readily apparent from the patent application, that polypeptides encoded by the claimed polynucleotides share more than 90% sequence identity over 222 amino acid residues with the SEQ ID NO:1 polypeptide. This is more than enough homology to demonstrate a reasonable probability that the utility of the SEQ ID NO:1 polypeptide can be imputed to the claimed polynucleotides (through the polypeptides they encode). It is well-known that the probability that two unrelated polypeptides share more than 40% sequence homology over 70 amino acid residues is exceedingly small. (Brenner et al., *supra*, Reference No. 1). Given homology in excess of 40% over many more than 70 amino acid residues, the probability that the polypeptide encoded for by the claimed polynucleotide is related to the SEQ ID NO:1 polypeptide is, accordingly, very high.

The Examiner must accept the Applicants' demonstration that the homology between the polypeptides encoded by the claimed invention and the SEQ ID NO:1 polypeptide demonstrates utility by a reasonable probability unless the Examiner can demonstrate through evidence or sound scientific reasoning that a person of ordinary skill in the art would doubt utility. See *In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has not provided sufficient evidence or sound scientific reasoning to the contrary.

While the Examiner has cited literature (Russell, J. Mol. Bio. 244:332-350, Skolnick et al. Trends in Biotech. 18(1):34-39, and Attwood [*sic*: Atwood] Science 290: 471-473, 29 October 2000) identifying some of the difficulties that may be involved in predicting protein function, none suggests that functional homology cannot be inferred by a reasonable probability in this case. Most important, none contradicts Brenner's basic rule that sequence homology in excess of 40% over 70 or more amino acid residues yields a high probability of functional homology as well. At most, these articles individually and together stand for the proposition that it is difficult to make predictions about function with certainty. The standard applicable in this case is not, however, proof to certainty, but rather proof to reasonable probability.

The Examiner contends that the use of sequence identity of the SEQ ID NO:2 polynucleotide to the claimed polynucleotide variants is insufficient to identify the function of the claimed protein, relying

neither evidence nor sound scientific reasoning to support the allegations that the claimed polynucleotides lacked adequate enablement. The Examiner made only the bare statement that "[s]ee the following publications that support this unpredictability as well as noting certain conserved sequences in limited specific cases." (Office Action, page 10.) The Examiner pointed to no specific part of any of the cited documents that supported the allegations of lack of enablement.

Furthermore, it is well known in the art that sequence similarity (measured by statistical scores as in Brenner) is predictive of similarity in functional activity. H. Hegyi and M. Gerstein ("The Relationship between Protein Structure and Function: a Comprehensive Survey with Application to the Yeast Genome," J. Mol. Biol. (1999) 288:147-164; Reference No. 8) state that "the proportion of homologues with different functions is around 10%. This shows that **there is a low chance that a single-domain protein, highly homologous to a known enzyme, has a different function.**" (Hegyi and Gerstein, Reference No. 8, page 159, column 1, emphasis added.) Furthermore, Hegyi and Gerstein in a second journal article (H. Hegyi and M. Gerstein, "Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins," Genome Research (2001) 11: 1632-1640; Reference No. 9) conclude that "the probability that two single-domain proteins that have the same superfamily structure have the same function (whether enzymatic or not) is about 2/3." (Hegyi and Gerstein, Reference No. 9, page 1635.) Hegyi and Gerstein also concluded that, for multi-domain proteins with "almost complete coverage with exactly the same type and number of superfamilies, following each other in the same order" "[t]he probability that the functions are the same in this case was 91%." (Hegyi and Gerstein, Reference No. 9, page 1636.) Hegyi and Gerstein (Reference No. 9, page 1632)) further note that

> Wilson et al. (2000) compared a large number of protein domains to one another in a pair-wise fashion with respect to similarities in sequence, structure, and function. Using a hybrid functional classification scheme merging the ENZYME and FlyBase systems (Gelbart et al. 1997; Bairoch 2000), they found that precise function is not conserved below 30–40% identity, although the broad functional class is usually preserved for sequence identities as low as 20–25%, given that the sequences have the same fold. Their survey also reinforced the previously established general exponential relationship [illegible] 1986).

As noted *supra* the polypeptides encoded by the claimed polynucleotides share more than 90% sequence identity with the SEQ ID NO:1 polypeptide, well above the thresholds described in the Hegyi and Gerstein Genome Research article (Reference No. 9) cited above. Therefore, there is a reasonable probability that the utility of the SEQ ID NO:1 polypeptide can be imputed to the claimed polynucleotides.

Applicants further submit that both the Revised Interim Utility Guidelines and the Revised Interim Utility Guidelines Training Materials support the use of sequence homology to known proteins to establish functional homology. The Revised Interim Utility Guidelines specifically state at page 1096, that the Examiner's decision to rebut Applicants assertion of utility:

> ---must be supported by a preponderance of all evidence of record. More specifically, when a patent application claiming a nucleic acid asserts a specific, substantial, and credible utility, and bases the assertion upon homology to existing nucleic acids or proteins having an accepted utility, the asserted utility must be accepted by the Examiner unless the Office has sufficient evidence or sound scientific reasoning to rebut such an assertion. "[A] 'rigorous correlation' need not be shown in order to establish practical utility; 'reasonable correlation' is sufficient". (emphasis added).

Clearly the PTO recognizes the well known use of sequence homology in the art to establish protein function. The Revised Interim Utility Guidelines Training Materials elaborate further on this matter in Example 10:DNA Fragment encoding a Full Open Reading Frame (ORF) at page 53, which recites a claim to a nucleic acid encoding a protein with 95% sequence identity to a known protein (a DNA ligase). The example clearly states that "there is no reason to doubt the assertion that [the claimed sequence] encodes a DNA ligase." Therefore the Revised Interim Utility Guidelines Training Materials indicate that a sequence similarity of less than 100% is deemed reasonably to support to one skilled in the art that two molecules could possess the same activity.

Moreover, In a recent Federal Circuit decision (Boehringer Ingelheim Vetmedica, Inc. v. Schering-Plough Corporation and Schering Corporation; CAFC 02-1026, -1027, February 21, 2003), the Court stated that "the uncontroversial fact that even a single nucleotide or amino acid substitution may drastically alter the function of a gene or protein is not evidence of anything at all. The

mere possibility that a single mutation could affect biological function cannot as a matter of law preclude an assertion of equivalence."

d. Objective evidence corroborates the utilities of the claimed invention

There is, in fact, no restriction on the kinds of evidence a Patent Examiner may consider in determining whether a "real-world" utility exists. Indeed, "real-world" evidence, such as evidence showing actual use or commercial success of the invention, can demonstrate conclusive proof of utility. *Raytheon v. Roper*, 220 USPQ2d 592 (Fed. Cir. 1983); *Nestle v. Eugene*, 55 F.2d 854, 856, 12 USPQ 335 (6th Cir. 1932). Indeed, proof that the invention is made, used or sold by any person or entity other than the patentee is conclusive proof of utility. *United States Steel Corp.* v. *Phillips Petroleum Co.*, 865 F.2d 1247, 1252, 9 USPQ2d 1461 (Fed. Cir. 1989).

Over the past several years, a vibrant market has developed for databases containing the sequences of all expressed genes (along with the polypeptide translations of those genes), in particular genes having medical and pharmaceutical significance such as the instant sequences. (Note that the value in these databases is enhanced by their completeness, but each sequence in them is independently valuable.) The databases sold by Applicants' assignee, Incyte, include exactly the kinds of information made possible by the claimed invention, such as tissue and disease associations. Incyte sells its database containing the sequences of the claimed polynucleotides and millions of other sequences throughout the scientific community, including to pharmaceutical companies who use the information to develop new pharmaceuticals.

Both Incyte's customers and the scientific community have acknowledged that Incyte's databases have proven to be valuable in, for example, the identification and development of drug candidates. As Incyte adds information to its databases, including the information that can be generated only as a result of Incyte's discovery of the claimed polynucleotides and its use of those polynucleotides on cDNA microarrays, the databases become even more powerful tools. Thus the claimed invention adds more than incremental benefit to the drug discovery and development process.

3. The Patent Examiner's Rejections Are Without Merit

Rather than responding to the evidence demonstrating utility, the Examiner attempts to dismiss it altogether by arguing that the disclosed and well-established uses for the claimed polynucleotides are not enabled. (Office Action at pages 9-10.) The Examiner is incorrect both as a matter of law and as a matter of fact.

a. The Precise Biological Role Or Function Of An Expressed Polynucleotide Is Not Required To Demonstrate Utility

The Patent Examiner's primary rejection of the claimed invention is based on the ground that, without information as to the precise "biological function" of the claimed invention, the claimed invention's utility is not sufficiently specific. According to the Examiner, it is not enough that a person of ordinary skill in the art could use and, in fact, would want to use the claimed invention either by itself or in a cDNA microarray to monitor the expression of genes for such applications as the evaluation of a drug's efficacy and toxicity. The Examiner would require, in addition, that the applicant provide a specific and substantial interpretation of the results generated in any given expression analysis.

It may be that specific and substantial interpretations and detailed information on biological function are necessary to satisfy the requirements for publication in some technical journals, but they are not necessary to satisfy the requirements for obtaining a United States patent. The relevant question is not, as the Examiner would have it, whether it is known how or why the invention works, *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999), but rather whether the invention provides an "identifiable benefit" in presently available form. *Juicy Whip Inc. v. Orange Bang Inc.*, 185 F.3d 1364, 1366 (Fed. Cir. 1999). If the benefit exists, and there is a substantial likelihood the invention provides the benefit, it is useful. There can be no doubt, particularly in view of the Bedilion Declaration (at, *e.g.*, ¶¶ 10 and 15, Bedilion), that the present invention meets this test.

The threshold for determining whether an invention produces an identifiable benefit is low. *Juicy Whip*, 185 F.3d at 1366. Only those utilities that are so nebulous that a person of ordinary skill

guidelines, so-called "throwaway" utilities that are not directed to a person of ordinary skill in the art at all, do not meet the statutory requirement of utility. Utility Examination Guidelines, 66 Fed. Reg. 1092 (Jan. 5, 2001).

Knowledge of the biological function or role of a biological molecule has never been required to show real-world benefit. In its most recent explanation of its own utility guidelines, the PTO acknowledged so much (66 F.R. at 1095):

> [T]he utility of a claimed DNA does not necessarily depend on the function of the encoded gene product. A claimed DNA may have specific and substantial utility because, *e.g.*, it hybridizes near a disease-associated gene or it has gene-regulating activity.

By implicitly requiring knowledge of biological function for any claimed nucleic acid, the Examiner has, contrary to law, elevated what is at most an evidentiary factor into an absolute requirement of utility. Rather than looking to the biological role or function of the claimed invention, the Examiner should have looked first to the benefits it is alleged to provide.

b. Membership in a Class of Useful Products Can Be Proof of Utility

Despite the uncontradicted evidence that the claimed polynucleotides encode polypeptides in the glutathione s-transferase family and the family of expressed polypeptides, the Examiner refused to impute the utility of the members of the glutathione s-transferase family and the family of expressed polypeptides to the polypeptides encoded by the claimed polynucleotides.

In order to demonstrate utility by membership in a class, the law requires only that the class not contain a substantial number of useless members. So long as the class does not contain a substantial number of useless members, there is sufficient likelihood that the claimed invention will have utility, and a rejection under 35 U.S.C. § 101 is improper. That is true regardless of how the claimed invention ultimately is used and whether or not the members of the class possess one utility or many. *See Brenner v. Manson*, 383 U.S. 519, 532 (1966); *Application of Kirk*, 376 F.2d 936, 943 (CCPA

Membership in a "general" class is insufficient to demonstrate utility only if the class contains a sufficient number of useless members such that a person of ordinary skill in the art could not impute utility by a substantial likelihood. There would be, in that case, a substantial likelihood that the claimed invention is one of the useless members of the class. In the few cases in which class membership did not prove utility by substantial likelihood, the classes did in fact include predominately useless members. *E.g.*, *Brenner* (man-made steroids); *Kirk* (same); *Natta* (man-made polyethylene polymers).

The Examiner addresses the polypeptides encoded by the claimed polynucleotides as if the general classes in which they are included are not the glutathione s-transferase family and the family of expressed polypeptides, but rather all polynucleotides or all polypeptides, including the vast majority of useless theoretical molecules not occurring in nature, and thus not pre-selected by nature to be useful. While these "general classes" may contain a substantial number of useless members, the glutathione s-transferase family and the family of expressed polypeptides do not. The glutathione s-transferase family and the family of expressed polypeptides are sufficiently specific to rule out any reasonable possibility that the polypeptides encoded by the claimed polynucleotides would not also be useful like the other members of the family.

Because the Examiner has not presented any evidence that the glutathione s-transferase family and the family of expressed polypeptides have any, let alone a substantial number, of useless members, the Examiner must conclude that there is a "substantial likelihood" that the polypeptides encoded by the claimed polynucleotides are useful. It follows that the claimed polynucleotides also are useful.

c. Because the uses of the claimed polynucleotides in toxicology testing, drug discovery, and disease diagnosis are practical uses beyond mere study of the invention itself, the claimed invention has substantial utility.

As used in toxicology testing, drug discovery, and disease diagnosis, the claimed invention has a beneficial use in research other than studying the claimed invention or its protein products. It is a tool, rather than an object, of research. The data generated in gene expression monitoring using the claimed

study properties of tissues, cells, and potential drug candidates and toxins. Without the claimed invention, the information regarding the properties of tissues, cells, drug candidates and toxins is less complete. [Bedilion Declaration at ¶ 15.]

The claimed invention has numerous additional uses as a research tool, each of which alone is a "substantial utility." These include diagnostic assays (e.g., page 34, line 18 through page 37, line 11) and chromosomal mapping (e.g., page 37, line 12 through page 38, line 10).

d. The Patent Examiner Failed to Demonstrate That a Person of Ordinary Skill in the Art Would Reasonably Doubt the Utility of the Claimed Invention

Based principally on citations to scientific literature identifying some of the difficulties involved in predicting protein function, the Examiner rejected the pending claims on the ground that the Applicants cannot impute utility to the claimed polynucleotides based on their 90% sequence similarity to the SEQ ID NO:2 polynucleotide. The Examiner's rejection is both incorrect as a matter of fact and as a matter of procedural law.

As demonstrated in § VI.B.2.c., *supra*, the literature cited by the Examiner is not inconsistent with the Applicants' proof of homology by a reasonable probability. It may show that Applicants cannot prove function by homology with **certainty**, but Applicants need not meet such a rigorous standard of proof. Under the applicable law, once the applicant demonstrates a *prima facie* case of homology, the Examiner must accept the assertion of utility to be true unless the Examiner comes forward with evidence showing a person of ordinary skill would doubt the asserted utility could be achieved by a reasonable probability. *See In re Brana*, 51 F.3d at 1566; *In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has not made such a showing and, as such, the Examiner's rejection should be withdrawn.

4. By Requiring the Patent Applicant to Assert a Particular or Unique Utility, the Patent Examination Utility Guidelines and Training Materials Applied by the Patent Examiner Misstate the Law

There is an additional, independent reason to withdraw the rejections: to the extent the rejections are based on Revised Interim Utility Examination Guidelines (64 FR 71427, December 21, 1999), the final Utility Examination Guidelines (66 FR 1092, January 5, 2001) and/or the Revised Interim Utility Guidelines Training Materials (USPTO Website www.uspto.gov, March 1, 2000), the Guidelines and Training Materials are themselves inconsistent with the law.

The Training Materials, which direct the Examiners regarding how to apply the Utility Guidelines, address the issue of specificity with reference to two kinds of asserted utilities: "specific" utilities which meet the statutory requirements, and "general" utilities which do not. The Training Materials define a "specific utility" as follows:

> A [specific utility] is *specific* to the subject matter claimed. This contrasts to *general* utility that would be applicable to the broad class of invention. For example, a claim to a polynucleotide whose use is disclosed simply as "gene probe" or "chromosome marker" would not be considered to be specific in the absence of a disclosure of a specific DNA target. Similarly, a general statement of diagnostic utility, such as diagnosing an unspecified disease, would ordinarily be insufficient absent a disclosure of what condition can be diagnosed.

The Training Materials distinguish between "specific" and "general" utilities by assessing whether the asserted utility is sufficiently "particular," *i.e.*, unique (Training Materials at p.52) as compared to the "broad class of invention." (In this regard, the Training Materials appear to parallel the view set forth in Stephen G. Kunin, Written Description Guidelines and Utility Guidelines, 82 J.P.T.O.S. 77, 97 (Feb. 2000) ("With regard to the issue of specific utility the question to ask is whether or not a utility set forth in the specification is *particular* to the claimed invention.")).

Such "unique" or "particular" utilities never have been required by the law. To meet the utility requirement, the invention need only be "practically useful," *Natta*, 480 F.2d 1 at 1397, and confer a "specific benefit" on the public. *Brenner*, 383 U.S. at 534. Thus, incredible "throwaway" utilities, such as trying to "patent a transgenic mouse by saying it makes great snake food," do not meet this standard. Karen Hall, Genomic Warfare, The American Lawyer 68 (June 2000) (quoting John Doll, Chief of the Biotech Section of USPTO).

that are unique to an invention. The law requires that the practical utility be "definite," not particular. *Montedison*, 664 F.2d at 375. Applicants are not aware of any court that has rejected an assertion of utility on the grounds that it is not "particular" or "unique" to the specific invention. Where courts have found utility to be too "general," it has been in those cases in which the asserted utility in the patent disclosure was not a practical use that conferred a specific benefit. That is, a person of ordinary skill in the art would have been left to guess as to how to benefit at all from the invention. In *Kirk*, for example, the CCPA held the assertion that a man-made steroid had "useful biological activity" was insufficient where there was no information in the specification as to how that biological activity could be practically used. *Kirk*, 376 F.2d at 941.

The fact that an invention can have a particular use does not provide a basis for requiring a particular use. *See Brana, supra* (disclosure describing a claimed antitumor compound as being homologous to an antitumor compound having activity against a "particular" type of cancer was determined to satisfy the specificity requirement). "Particularity" is not and never has been the *sine qua non* of utility; it is, at most, one of many factors to be considered.

As described *supra*, broad classes of inventions can satisfy the utility requirement so long as a person of ordinary skill in the art would understand how to achieve a practical benefit from knowledge of the class. Only classes that encompass a significant portion of nonuseful members would fail to meet the utility requirement. *Supra* § VI.B.3.b. (*Montedison*, 664 F.2d at 374-75).

The Training Materials fail to distinguish between broad classes that convey information of practical utility and those that do not, lumping all of them into the latter, unpatentable category of "general" utilities. As a result, the Training Materials paint with too broad a brush. Rigorously applied, they would render unpatentable whole categories of inventions that heretofore have been considered to be patentable and that have indisputably benefitted the public, including the claimed invention. See *supra* §VI.B.3.b. Thus the Training Materials cannot be applied consistently with the law.

VII. Rejection of Claims 2-4 Under 35 U.S.C. § 112, second paragraph

> Claim 2 recites the limitation "an isolated polynucleotide encoding a polypeptide"; and claim 3 recites the limitation "a recombinant polynucleotide" however claim 1 does not recite a polynucleotide. There is insufficient antecedent basis for this limitation in the claim. (Office Action, page 11.)

In order to expedite prosecution, Claim 2 has been amended to an independent claim. Claims 3 and 4 depend from independent Claim 2. For at least the above reasons, Applicants respectfully request that the Examiner withdraw the indefiniteness rejection.

VIII. Rejection of Claim 8 Under the Judicially Created Doctrine of Obviousness-type Double Patenting

The Examiner rejected Claim 8 under the judicially created doctrine of obviousness-type double patenting as being unpatentable over Claims 3-4 of U.S. Patent No. 5,817,497 (Office Action, pages 11-12). Applicants request that the requirement for submission of a Terminal Disclaimer be held in abeyance until such time as there is an indication of allowable subject matter.

IX. Rejection of Claims 8-9 Under 35 U.S.C. § 102(b) as Being Anticipated by Hillier et al.

The Examiner rejected Claims 8-9 under 35 U.S.C. § 102(b) as being anticipated by Hillier et al. (Accession Number H27975). The Examiner alleged that "Hillier et al., teach a human cDNA clone that is similar to mouse glutathione transferase GST," that "[t]here are stretches of nucleic acids within the sequence that are complementary to SEQ ID NO:2," and that "the sequence of Hillier et al., recites a polynucleotide comprising at least 60 contiguous nucleic acids." (Office Action, page 12.)

Claim 9 is canceled, and Claim 8 is amended as follows:

> An isolated polynucleotide comprising a sequence selected from the group consisting of:
>
> a) a polynucleotide sequence of SEQ ID NO:2,
>
> b) a naturally-occurring polynucleotide sequence having at least 90% sequence identity to the sequence of SEQ ID NO:2, over the entire length of SEQ ID NO:2,
>
> c) a polynucleotide sequence completely complementary to a),
>
> d) a polynucleotide sequence completely complementary to b) and
>
> [illegible]

The Hillier et al. document does not teach a polynucleotide of Claim 8. For at least the above reasons, Applicants respectfully request that the Examiner withdraw the novelty rejection.

X. Rejection of Claims 2-4 Under 35 U.S.C. § 103(a) as Being Unpatentable Over Hillier et al. in View of Simula et al.

The Examiner rejected Claims 2-4 under 35 U.S.C. § 103(a) as being unpatentable over Hillier et al. (Accession Number H27975) in view of Simula et al. The Examiner stated that "Hillier et al., has been discussed above, however Hillier et al., does not disclose recombinant transformed cells comprising promoters," that "Simula et al., developed *Salmonella typhirumurium* [*sic: typhimurium*] strains that express human glutathione S-transferase (GST) (abstract). The GST was expressed using regulatable *tac* promoter expression systems (abstract)," and that "it would have been prima facie obvious to modify the cell transformed with a recombinant polynucleotide as taught by Simula et al., with the polynucleotide of Hillier et al." (Office Action, page 13.)

Claim 2 is amended as follows:

> 2. An isolated polynucleotide encoding a polypeptide comprising an amino acid sequence selected from the group consisting of:
> a) an amino acid sequence of SEQ ID NO:1, and
> b) a naturally-occurring amino acid sequence having at least 90% sequence identity to the sequence of SEQ ID NO:1 over the entire length of SEQ ID NO:1.

Neither Hillier et al. document nor the Simula et al. document teaches or suggests a polynucleotide encoding a polypeptide comprising an amino acid sequence selected from the group consisting of a) an amino acid sequence of SEQ ID NO:1, and b) a naturally-occurring amino acid sequence having at least 90% sequence identity to the sequence of SEQ ID NO:1 over the entire length of SEQ ID NO:1. Therefore, neither the Hillier et al. document nor the Simula et al. document, either alone or in combination, renders obvious Claims 2-4. For at least the above reasons, Applicants respectfully request that the Examiner withdraw the obviousness rejection

CONCLUSION

In light of the above amendments and remarks, Applicants submit that the present application is fully in condition for allowance, and request that the Examiner withdraw the outstanding objections and rejections. Early notice to that effect is earnestly solicited.

If the Examiner contemplates other action, or if a telephone conference would expedite allowance of the claims, Applicants invite the Examiner to contact Applicants' Agent at (650) 845-4646.

Applicants believe that no fee is due with this communication. However, if the USPTO determines that a fee is due, the Commissioner is hereby authorized to charge Deposit Account No. **09-0108.**

Respectfully submitted,

INCYTE GENOMICS, INC.

Date: March 13, 2003

Susan K. Sather

Susan K. Sather
Reg. No. 44,316
Direct Dial Telephone: (650) 845-4646

3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555
Fax: (650) 849-8886

VERSION WITH MARKINGS TO SHOW CHANGES MADE

IN THE SPECIFICATION:

Paragraph beginning at page 1, line 1, has been amended as follows:

This application is a divisional application of U.S. application serial number 09/309,320, filed May 11, 1999, issued June 19, 2001, as U.S. Patent No. 6,248,325, which is a divisional of U.S. application serial number 09/096,571, filed June 12, 1998, issued November 2, 1999, as U.S. Patent No. 5,976,528, which is a divisional application of U.S. application serial number 08/756,771, filed November 26, 1996, issued October 6, 1998, as U.S. Patent No. 5,817,497. U.S. application serial numbers 09/309,320, 09/096,571, and 08/756,771 are hereby expressly incorporated by reference.

IN THE CLAIMS:

Claim 9 has been canceled.

Claims 2 and 8 have been amended as follows:

2. (Once Amended) An isolated polynucleotide encoding a polypeptide [of claim 1] comprising an amino acid sequence selected from the group consisting of:

a) an amino acid sequence of SEQ ID NO:1, and

b) a naturally-occurring amino acid sequence having at least 90% sequence identity to the sequence of SEQ ID NO:1 over the entire length of SEQ ID NO:1.

8. (Once Amended) An isolated polynucleotide comprising a sequence selected from the group consisting of:

a) a polynucleotide sequence of SEQ ID NO:2,

b) a naturally-occurring polynucleotide sequence having at least 90% sequence identity

d) a polynucleotide sequence completely complementary to b) and

e) a ribonucleotide equivalent of a)-d).

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom, and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins), however, these complementary goals are linked such that increasing one [illegible]

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being [illegible]

[illegible]

[illegible]

‡To whom reprint requests should be addressed. e-mail: [illegible] hyper.stanford.edu

superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ≈0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of SCOP may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties −12/−1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

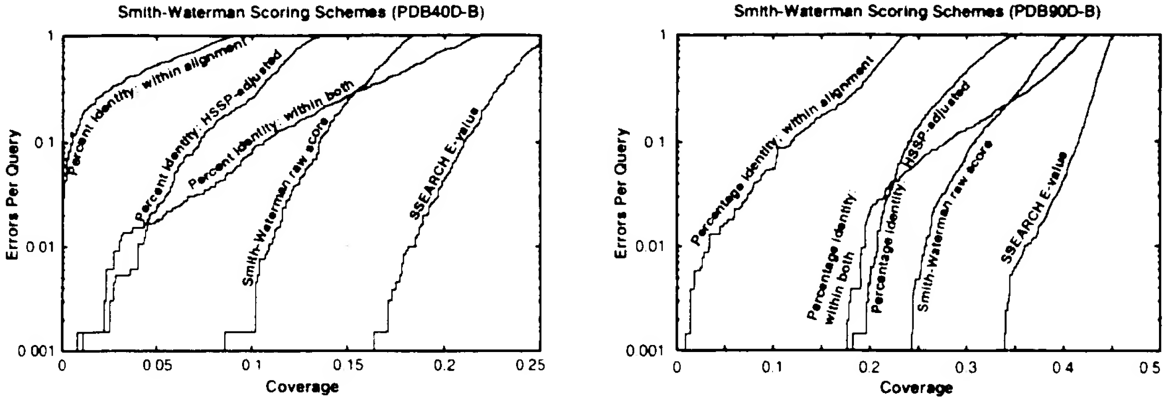


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the *x* axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The *y* axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The *y* axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

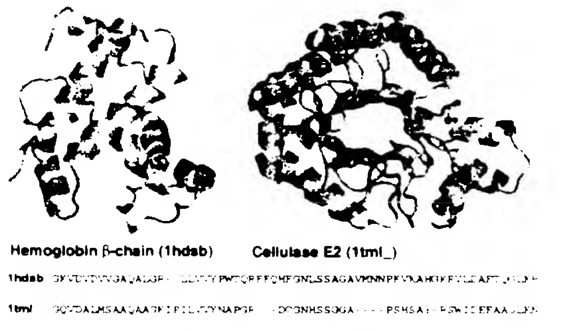

Hemoglobin β-chain (1hdsb)
Cellulase E2 (1tml_)
1hdsb GKVDVDVVGAQALGR - - LLVVYPWTQRFFQHFGNLSSAGAVMNNPKVKAHGKRVLDAFT
1tml GQVDALMSAAQAAGKIPILVVYNAPGR - - DCGNHSSGGA - - - - PSHSAY - RSWIDEFAA

FIG. 2. Unrelated proteins with high percentage identity. Hem [illegible] score of 85 nor the E-value of [illegible] is significant. Proteins rendered by RASMOL (40).

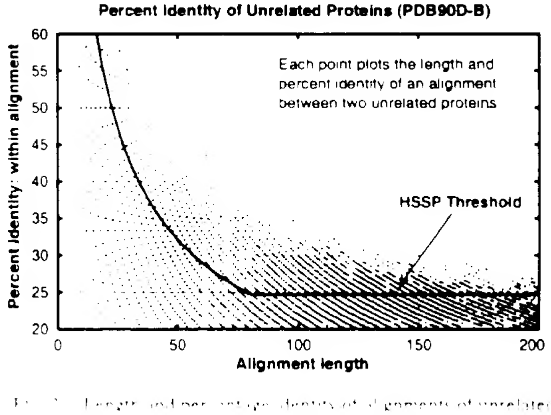


FIG. 3. Length and percentage identity of alignments of unrelated [illegible] The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

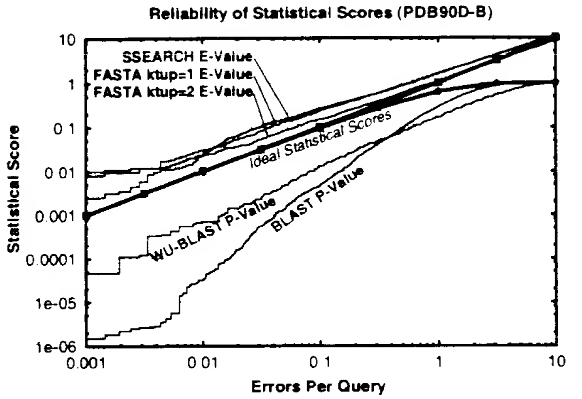


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith–Waterman" score, which is the measure optimized by the Smith–Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith–Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

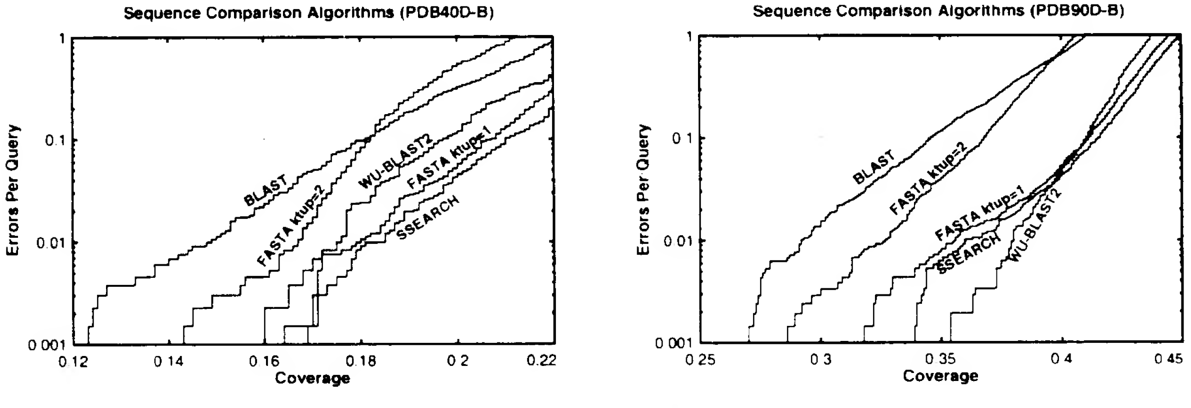


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (*A*) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (*B*) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

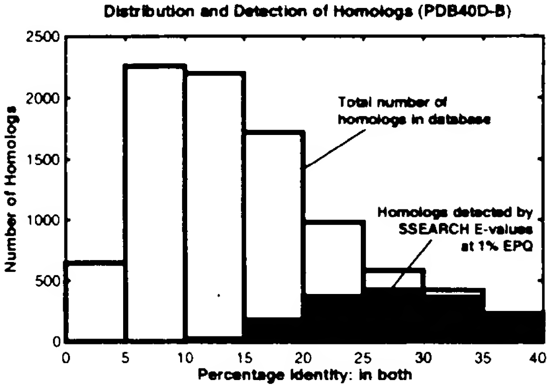


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pariwise sequence comparison to detect them.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	< 0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	[illegible]		

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at http://sss.stanford.edu/sss/.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266**, 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266**, 635–643.
6. Pearson, W. R. (1991) *Genomics* **11**, 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4**, 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266**, 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249**, 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17**, 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24**, 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24**, 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233**, 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1**, 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1**, 77–78.
21. Arratia, R., Gordon, L. & M. W. (1986) *Ann. Stat.* **14**, 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6**, 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266**, 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266**, 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9**, 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13**, 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7**, 369–376.
32. Orengo, C., Michie, A., Jones S, Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5**, 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39**, 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20**, 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16**, 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4**, 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131**, 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32**, 9906–9916.
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20**, 374–376.

XENOBIOTICA, 1999, VOL. 29, NO. 7, 655–691

Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential

JOHN C. ROCKETT†, DAVID J. ESDAILE‡
and G. GORDON GIBSON*

Molecular Toxicology Laboratory, School of Biological Sciences, University of Surrey,
Guildford, Surrey, GU2 5XH, UK

Received January 8, 1999

1. An important feature of the work of many molecular biologists is identifying which genes are switched on and off in a cell under different environmental conditions or subsequent to xenobiotic challenge. Such information has many uses, including the deciphering of molecular pathways and facilitating the development of new experimental and diagnostic procedures. However, the student of gene hunting should be forgiven for perhaps becoming confused by the mountain of information available as there appears to be almost as many methods of discovering differentially expressed genes as there are research groups using the technique.

2. The aim of this review was to clarify the main methods of differential gene expression analysis and the mechanistic principles underlying them. Also included is a discussion on some of the practical aspects of using this technique. Emphasis is placed on the so-called 'open' systems, which require no prior knowledge of the genes contained within the study model. Whilst these will eventually be replaced by 'closed' systems in the study of human, mouse and other commonly studied laboratory animals, they will remain a powerful tool for those examining less fashionable models.

3. The use of suppression-PCR subtractive hybridization is exemplified in the identification of up- and down-regulated genes in rat liver following exposure to phenobarbital, a well-known inducer of the drug metabolizing enzymes.

4. Differential gene display provides a coherent platform for building libraries and microchip arrays of 'gene fingerprints' characteristic of known enzyme inducers and xenobiotic toxicants, which may be interrogated subsequently for the identification and characterization of xenobiotics of unknown biological properties.

Introduction

It is now apparent that the development of almost all cancers and many non-neoplastic diseases are accompanied by altered gene expression in the affected cells compared to their normal state (Hunter 1991, Wynford-Thomas 1991, Vogelstein and Kinzler 1993, Semenza 1994, Cassidy 1995, Kleinjan and Van Heyningen 1998). Such changes also occur in response to external stimuli such as pathogenic micro-organisms (Rohn *et al.* 1996, Singh *et al.* 1997, Griffin and Krishna 1998, Lunney 1998) and xenobiotics (Sewall *et al.* 1995, Dogra *et al.* 1998, Ramana and Kohli 1998), as well as during the development of undifferentiated cells (Hecht 1998, Rudin and Thompson 1998, Schneider-Maunoury *et al.* 1998). The potential medical and therapeutic benefits of understanding the molecular changes which occur in any given cell in progressing from the normal to the 'altered' state are enormous. Such profiling essentially provides a 'fingerprint' of each step of a

* Author for correspondence, e-mail g.gibson@surrey.ac.uk

http www.tandf.co.uk JNLS xen.htm   http www.taylorandfrancis.com JNLS xen.htm

UK

area network to the online edition.

ridge University Press The online

rammes and details on advertising

Avenue Elmont New York 11003
New York 11003

he pound sterling price applies All
Canada, Mexico, India, Japan and
be made by sterling cheque, dollar

Post Bag No 8, Saket, New Delhi

PA 19106
4 5PR

ion may be reproduced, stored,
(from Taylor & Francis Limited,

cell's development or response and should help in the elucidation of specific and sensitive biomarkers representing, for example, different types of cancer or previous exposure to certain classes of chemicals that are enzyme inducers.

In drug metabolism, many of the xenobiotic-metabolizing enzymes (including the well-characterized isoforms of cytochrome P450) are inducible by drugs and chemicals in man (Pelkonen *et al.* 1998), predominantly involving transcriptional activation of not only the cognate cytochrome P450 genes, but additional cellular proteins which may be crucial to the phenomenon of induction. Accordingly, the development of methodology to identify and assess the full complement of genes that are either up- or down-regulated by inducers are crucial in the development of knowledge to understand the precise molecular mechanisms of enzyme induction and how this relates to drug action. Similarly, in the field of chemical-induced toxicity, it is now becoming increasingly obvious that most adverse reactions to drugs and chemicals are the result of multiple gene regulation, some of which are causal and some of which are casually-related to the toxicological phenomenon *per se*. This observation has led to an upsurge in interest in gene-profiling technologies which differentiate between the control and toxin-treated gene pools in target tissues and is, therefore, of value in rationalizing the molecular mechanisms of xenobiotic-induced toxicity. Knowledge of toxin-dependent gene regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. For example, if the gene profile in response to say a testicular toxin that has been well-characterized *in vivo* could be determined in the testis, then this profile would be representative of all new drug candidates which act via this specific molecular mechanism of toxicity, thereby providing a useful and coherent approach to the early detection of such toxicants. Whereas it would be informative to know the identity and functionality of all genes up/down regulated by such toxicants, this would appear a longer term goal, as the majority of human genes have not yet been sequenced, far less their functionality determined. However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well-characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. Such approaches are beginning to gain momentum, in that several biotechnology companies are commercially producing 'gene chips' or 'gene arrays' that may be interrogated for toxicity assessment of xenobiotics. These chips consist of hundreds/thousands of genes, some of which are degenerate in the sense that not all of the genes are mechanistically-related to any one toxicological phenomenon. Whereas these chips are useful in broad-spectrum screening, they are maturing at a substantial rate, in that gene arrays are now becoming more specific, e.g. chips for the identification of changes in growth factor families that contribute to the aetiology and development of chemically-induced neoplasias.

Although documenting and explaining these genetic changes presents a formidable obstacle to understanding the different mechanisms of development and disease progression, the technology is now available to begin attempting this difficult challenge. Indeed, several 'differential expression analysis' methods have been developed which facilitate the identification of gene products that demonstrate

ation of specific and
of cancer or previous
:ers.
enzymes (including
ucible by drugs and
lving transcriptional
it additional cellular
on. Accordingly, the
omplement of genes
n the development of
of enzyme induction
of chemical-induced
adverse reactions to
n, some of which are
ical phenomenon *per*
profiling technologies
pools in target tissues
inisms of xenobiotic-
on in target tissues is
'n generated in the
identification of toxic
ess and contributing
ole, if the gene profile
rized *in vivo* could be
ative of all new drug
of toxicity, thereby
on of such toxicants.
ctionality of all genes
iger term goal, as the
ss their functionality
ds a *pattern* of gene
itched to that of well-
e *in vivo* similarities
a platform for more
beginning to gain
mercially producing
oxicity assessment of
es, some of which are
tically-related to any
il in broad-spectrum
gene arrays are now
nges in growth factor
f chemically-induced

changes presents a
s of development and
empting this difficult

altered expression in cells of one population compared to another. These methods have been used to identify differential gene expression in many situations, including invading pathogenic microbes (Zhao *et al.* 1998), in cells responding to extracellular and intracellular microbial invasion (Duguid and Dinauer 1990, Ragno *et al.* 1997, Maldarelli *et al.* 1998), in chemically treated cells (Syed *et al.* 1997, Rockett *et al.* 1999), neoplastic cells (Liang *et al.* 1992, Chang and Terzaghi-Howe 1998), activated cells (Gurskaya *et al.* 1996, Wan *et al.* 1996), differentiated cells (Hara *et al.* 1991, Guimaraes *et al.* 1995a, b), and different cell types (Davis *et al.* 1984, Hedrick *et al.* 1984, Xhu *et al.* 1998). Although differential expression analysis technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases,'absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

The field of differential expression analysis is a large and complex one, with many techniques available to the potential user. These can be categorized into several methodological approaches, including:

(1) Differential screening,
(2) Subtractive hybridization (SH) (includes methods such as chemical cross-linking subtraction—CCLS, suppression-PCR subtractive hybridization—SSH, and representational difference analysis—RDA),
(3) Differential display (DD),
(4) Restriction endonuclease facilitated analysis (including serial analysis of gene expression—SAGE—and gene expression fingerprinting—GEF),
(5) Gene expression arrays, and
(6) Expressed sequence tag (EST) analysis.

The above approaches have been used successfully to isolate differentially expressed genes in different model systems. However, each method has its own subtle (and sometimes not so subtle) characteristics which incur various advantages and disadvantages. Accordingly, it is the purpose of this review to clarify the mechanistic principles underlying the main differential expression methods and to highlight some of the broader considerations and implications of this very powerful and increasingly popular technique. Specifically, we will concentrate on the so-called 'open' systems, namely those which do not require any knowledge of gene sequences and, therefore, are useful for isolating unknown genes. Two 'closed' systems (those utilising previously identified gene sequences), EST analysis and the use of DNA arrays, will also be considered briefly for completeness. Whilst emphasis will often be placed on suppression PCR subtractive hybridization (SSH, the approach employed in this laboratory), it is the aim of the authors to highlight, wherever possible, those areas of common interest to those who use, or intend to use, differential gene expression analysis.

Differential cDNA library screening (DS)

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years. One of the original hybridization', which was used to isolate galactose-inducible DNA sequences from yeast. The theory is simple: a genomic DNA library is prepared from normal, unstimulated cells of the test organism/tissue and multiple filter replicas are prepared. These replica blots are probed with radioactively (or otherwise) labelled complex cDNA probes prepared from the control and test cell mRNA populations. Those mRNAs which are differentially expressed in the treated cell population will show a positive signal only on the filter probed with cDNA from the treated cells. Furthermore, labelled cDNA from different test conditions can be used to probe multiple blots, thereby enabling the identification of mRNAs which are only up-regulated under certain conditions. For example, St John and Davis (1979) screened replica filters with acetate-, glucose- and galactose-derived probes in order to obtain genes induced specifically by galactose metabolism. Although groundbreaking in its time this method is now considered insensitive and time-consuming, as up to 2 months are required to complete the identification of genes which are differentially expressed in the test population. In addition, there is no convenient way to check that the procedure has worked until the whole process has been completed.

Subtractive Hybridization (SH)

The developing concept of differential gene expression and the success of early approaches such as that described by St John and Davis (1979) soon gave rise to a search for more convenient methods of analysis. One of the first to be developed was SH, numerous variations of which have since been reported (see below). In general, this approach involves hybridization of mRNA/cDNA from one population (tester) to excess mRNA/cDNA from another (driver), followed by separation of the unhybridized tester fraction (differentially expressed) from the hybridized common sequences. This step has been achieved physically, chemically and through the use of selective polymerase chain reaction (PCR) techniques.

Physical separation

Original subtractive hybridization technology involved the physical separation of hybridized common species from unique single stranded species. Several methods of achieving this have been described, including hydroxyapatite chromatography (Sargent and Dawid 1983), avidin-biotin technology (Duguid and Dinauer 1990) and oligodT-latex separation (Hara *et al.* 1991). In the first approach, common mRNA species are removed by cDNA (from test cells)-mRNA (from control cells) subtractive hybridization followed by hydroxyapatite chromatography, as hydroxyapatite specifically adsorbs the cDNA-mRNA hybrids. The unabsorbed cDNA is then used either for the construction of a cDNA library of differentially expressed genes (Sargent and Dawid 1983, Schneider *et al.* 1988) or directly as a probe to screen a preselected library (Zimmerman *et al.* 1980, Davis *et al.* 1984, Hedrick *et al.* 1984). A schematic diagram of the procedure is shown in figure 1.

Less rigorous physical separation procedures coupled with sensitivity enhancing PCR steps were later developed as a means to overcome some of the problems encountered with the hydroxyapatite procedure. For example, Daguid and Dinauer (1990) described a method of subtraction utilizing biotin-affinity systems as a means to remove hybridized common sequences. In this process, both the control and tester mRNA populations are first converted to cDNA and an adaptor ('oligovector',

)NA sequences from
epared from normal,
le filter replicas are
or otherwise) labelled
mRNA populations.
d cell population will
rom the treated cells.
can be used to probe
s which are only up-
Davis (1979) screened
bes in order to obtain
groundbreaking in its
nsuming, as up to 2
hich are differentially
venient way to check
en completed.

d the success of early
9) soon gave rise to a
t to be developed was
ee below). In general,
ne population (tester)
by separation of the
hybridized common
and through the use

e physical separation
cies. Several methods
tite chromatography
a and Dinauer 1990)
approach, common
A (from control cells)
ography, as hydroxy-
unabsorbed cDNA is
fferentially expressed
directly as a probe to
l. 1984, Hedrick *et al.*
re 1.
sensitivity enhancing
ome of the problems
Daguid and Dinauer

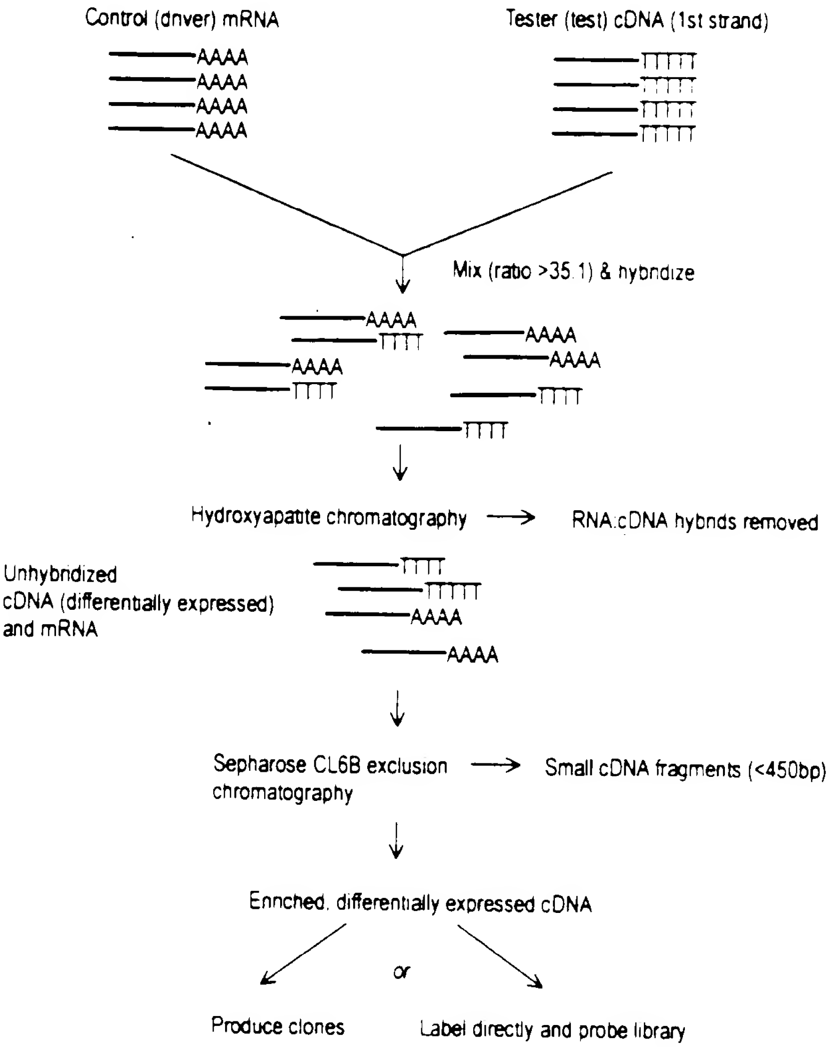

Control (driver) mRNA
Tester (test) cDNA (1st strand)
AAAA
TTTTT
Mix (ratio >35:1) & hybridize
AAAA
TTTT
Hydroxyapatite chromatography
RNA:cDNA hybrids removed
Unhybridized
cDNA (differentially expressed)
and mRNA
Sepharose CL6B exclusion
chromatography
Small cDNA fragments (<450bp)
Enriched, differentially expressed cDNA
or
Produce clones
Label directly and probe library

Figure 1. The hydroxyapatite method of subtractive hybridization. cDNA derived from the treated/altered (tester) population is mixed with a large excess of mRNA from the control (driver) population. Following hybridization, mRNA-cDNA hybrids are removed by hydroxyapatite chromatography. The only cDNAs which remain are those which are differentially expressed in the treated/altered population. In order to facilitate the recovery of full length clones, small cDNA fragments are removed by exclusion chromatography. The remaining cDNAs are then cloned into a vector for sequencing, or labelled and used directly to probe a library, as described by Sargent and Dawid (1983)

containing a restriction site) ligated to both sides. Both populations are then amplified by PCR, but the driver cDNA population is subsequently digested with the adaptor-containing restriction endonuclease. This serves to cleave the oligo-vector and reduce the amplification potential of the control population. The digested control population is then biotinylated and an excess mixed with tester cDNA. Following denaturation and hybridization, the mix is applied to a biocytin column

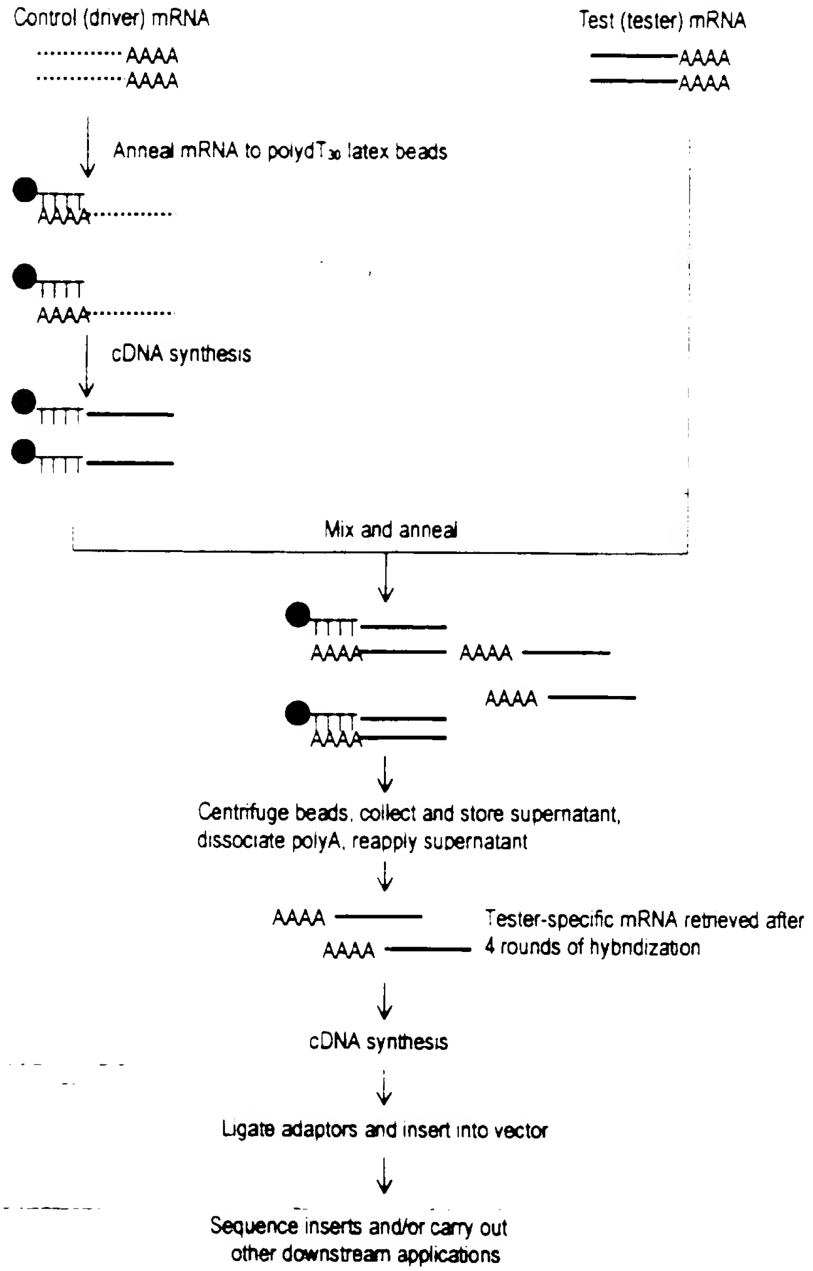
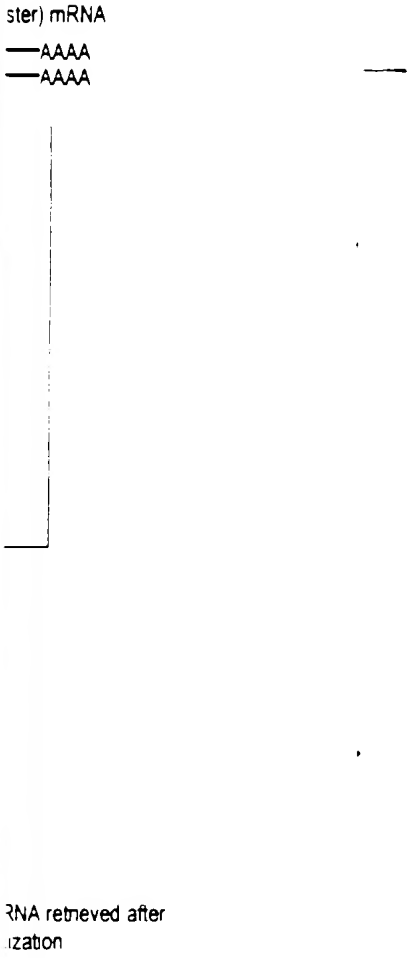

Control (driver) mRNA
AAAA
AAAA
Test (tester) mRNA
AAAA
AAAA
Anneal mRNA to polydT$_{30}$ latex beads
TTTT
AAAA
TTTT
AAAA
cDNA synthesis
TTTT
TTTT
Mix and anneal
TTTT
AAAA
AAAA
AAAA
TTTT
AAAA
Centrifuge beads, collect and store supernatant, dissociate polyA, reapply supernatant
AAAA
AAAA
Tester-specific mRNA retrieved after 4 rounds of hybridization
cDNA synthesis
Ligate adaptors and insert into vector
Sequence inserts and/or carry out other downstream applications

Figure 2. The use of oligodT$_{30}$ latex to perform subtractive hybridization. mRNA extracted from the control (driver) population is converted to anchored cDNA using polydT oligonucleotides attached to latex beads. mRNA from the treated/altered (tester) population is repeatedly hybridized against an excess of the anchored driver cDNA. The final population of mRNA is tester specific and can be converted into cDNA for cloning and other downstream applications, as described by Hara *et al.* (1991).



mRNA extracted from the
; polydT oligonucleotides
population is repeatedly
al population of mRNA is
ownstream applications, as

control cDNA. In order to further enrich those species differentially expressed in the tester cDNA, the subtracted tester population is amplified by PCR following every second subtraction cycle. After six cycles of subtraction (three reamplification steps) the reaction mix is ligated into a vector for further analysis.

In a slightly different approach, Hara *et al.* (1991) utilized a method whereby oligo(dT_{30}) primers attached to a latex substrate are used to first capture mRNA extracted from the control population. Following 1st strand cDNA synthesis, the RNA strand of the heteroduplexes is removed by heat denaturation and centrifugation (the cDNA-oligotex-dT_{30} forms a pellet and the supernatant is removed). A quantity of tester mRNA is then repeatedly hybridized to the immobilized control (driver) cDNA (which is present in 20-fold excess). After several rounds of hybridization the only mRNA molecules left in the tester mRNA population are those which are not found in the driver cDNA-oligotex-dT_{30} population. These tester-specific mRNA species are then converted to cDNA and, following the addition of adaptor sequences, amplified by PCR. The PCR products are then ligated into a vector for further analysis using restriction sites incorporated into the PCR primers. A schematic illustration of this subtraction process is shown in figure 2.

However, all these methods utilising physical separation have been described as inefficient due to the requirement for large starting amounts of mRNA, significant loss of material during the separation process and a need for several rounds of hybridization. Hence, new methods of differential expression analysis have recently been designed to eliminate these problems.

Chemical Cross-Linking Subtraction (CCLS)

In this technique, originally described by Hampson *et al.* (1992), driver mRNA is mixed with tester cDNA (1st strand only) in a ratio of > 20:1. The common sequences form cDNA:mRNA hybrids, leaving the tester specific species as single stranded cDNA. Instead of physically separating these hybrids, they are inactivated chemically using 2,5 diaziridinyl-1,4-benzoquinone (DZQ). Labelled probes are then synthesized from the remaining single stranded cDNA species (unreacted mRNA species remaining from the driver are not converted into probe material due to specificity of Sequenase T7 DNA polymerase used to make the probe) and used to screen a cDNA library made from the tester cell population. A schematic diagram of the system is shown in figure 3.

It has been shown that the differentially expressed sequences can be enriched at least 300-fold with one round of subtraction (Hampson *et al.* 1992), and that the technique should allow isolation of cDNAs derived from transcripts that are present at less than 50 copies per cell. This equates to genes at the low end of intermediate abundance (see table 1). The main advantages of the CCLS approach are that it is rapid, technically simple and also produces fewer false positives than other differential expression analysis methods. However, like the physical separation protocols, a major drawback with CCLS is the large amount of starting material required (at least 10 μg RNA). Consequently, the technique has recently been refined so that a renewable source of RNA can be generated. The degenerate random

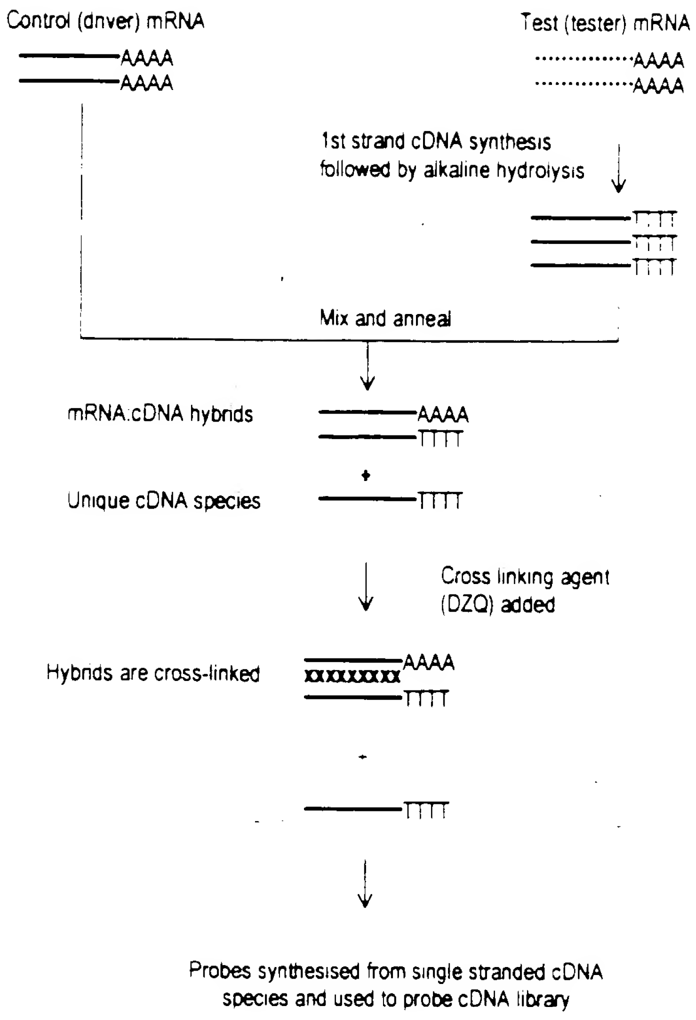


Figure 3. Chemical cross-linking subtraction. Excess driver mRNA is mixed with 1st strand tester cDNA. The common sequences form mRNA:cDNA hybrids which are cross linked with 2,5 diaziridinyl-1,4-benzoquinone (DZQ) and the remaining cDNA sequences are differentially expressed in the tester population. Probes are made from these sequences using Sequenase 2.0 DNA polymerase, which lacks reverse transcriptase activity and, therefore, does not react with the remaining mRNA molecules from the driver. The labelled probes are then used to screen a cDNA library for clones of differentially expressed sequences. Adapted from Walter *et al.* (1996), with permission.

Table 1. The abundance of mRNA species and classes in a typical mammalian cell.

mRNA class	Copies of each species/cell	No. of mRNA species in class	Mean % of each species in class	Mean mass (ng) of each species/μg total RNA
Abundant	12 000	4	3.3	1.65
Intermediate	300	500	0.08	0.04
Rare	15	11 000	0.004	0.002

Modified from Bertioli *et al.* (1995).

at the 5′ end, the final pool of random cDNA fragments is a PCR-renewable cDNA population which is representative of the expressed gene pool and can be used to synthesize sense RNA for use as driver material. Furthermore, if the final pool of random cDNA fragments is reamplified using biotinylated T7 primer and random hexamer, the product can be captured with streptavidin beads and the antisense strand eluted for use as tester. Since both target and driver can be generated from the same DROP product, subtraction can be performed in both directions (i.e. for up- and down-regulated species) between two different DROP products.

Representational Difference Analysis (RDA)

RDA of cDNA (Hubank and Schatz 1994) is an extension of the technique originally applied to genomic DNA as a means of identifying differences between two complex genomes (Lisitsyn *et al.* 1993). It is a process of subtraction and amplification involving subtractive hybridization of the tester in the presence of excess driver. Sequences in the tester that have homologues in the driver are rendered unamplifiable, whereas those genes expressed only in the tester retain the ability to be amplified by PCR. The procedure is shown schematically in figure 4.

In essence, the driver and tester mRNA populations are first converted to cDNA and amplified by PCR following the ligation of an adaptor. The adaptors are then removed from both populations and a new (different) adaptor ligated to the amplified tester population only. Driver and tester populations are next melted and hybridized together in a ratio of 100:1. Following hybridization, only tester:tester homohybrids have 5′ adaptors at each end of the DNA duplex and can, thus, be filled in at both 3′ ends. Hence, only these molecules are amplified exponentially during the subsequent PCR step. Although tester:driver heterohybrids are present, they only amplify in a linear fashion, since the strand derived from the driver has no adaptor to which the primer can bind. Driver:driver heterohybrids have no adaptors and, therefore, are not amplified. Single stranded molecules are digested with mung bean nuclease before a further PCR-enrichment of the tester:tester homohybrids. The adaptors on the amplified tester population are then replaced and the whole process repeated a further two or three times using an increasing excess of driver (Hubank and Shatz used a tester:driver ratio of 1:400, 1:80000 and 1:800000 for the second, third and fourth hybridizations, respectively). Different adaptors are ligated to the tester between successive rounds of hybridization and amplification to prevent the accumulation of PCR products that might interfere with subsequent amplifications. The final display is a series of differentially expressed gene products easily observable on an ethidium bromide gel.

The main advantages of RDA are that it offers a reproducible and sensitive approach to the analysis of differentially expressed genes. Hubank and Schatz (1994) reported that they were able to isolate genes that were differentially expressed in substantially less than 1 % of the cells from which the tester is derived. Perhaps the main drawback is that multiple rounds of ligation, hybridization, amplifiation and digestion are required. The procedure is, therefore, lengthier than many other differential display approaches and provides more opportunity for operator-induced error to occur. Although the generation of false positives has been noted, this has been solved to some degree by O'Neill and Sinclair (1997) through the use of HPLC-

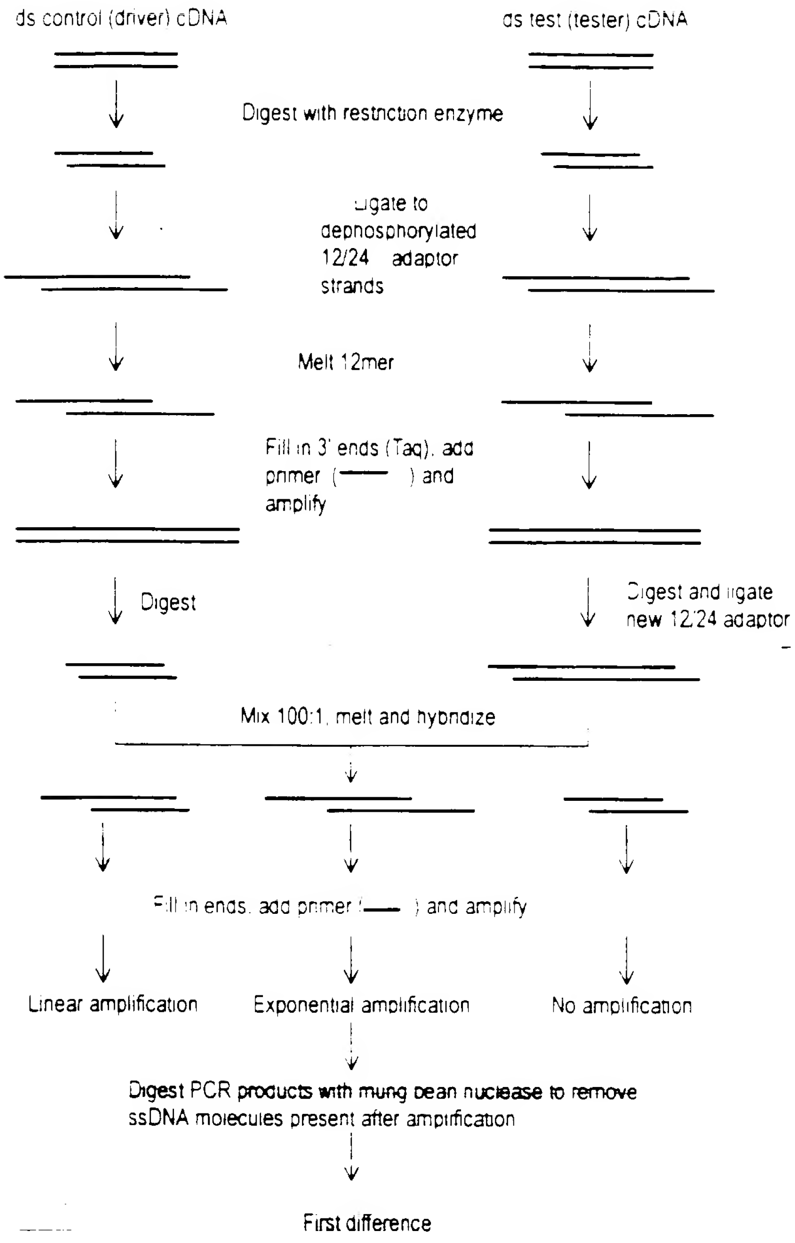

ds control (driver) cDNA
ds test (tester) cDNA
Digest with restriction enzyme
Ligate to dephosphorylated 12/24 adaptor strands
Melt 12mer
Fill in 3' ends (Taq), add primer ( ) and amplify
Digest
Digest and ligate new 12/24 adaptor
Mix 100:1, melt and hybridize
Fill in ends, add primer ( ) and amplify
Linear amplification
Exponential amplification
No amplification
Digest PCR products with mung bean nuclease to remove ssDNA molecules present after amplification
First difference

Figure 4. The representational difference analysis (RDA) technique. Driver and tester cDNA are digested with a 4-cutter restriction enzyme such as *Dpn*II. The 1st set of 12/24 adaptor strands (oligonucleotides) are ligated to each other and the digested cDNA products. The 12mer is subsequently melted away and the 3'ends filled in using Taq DNA polymerase. Each cDNA population is then amplified using PCR, following which the 1st set of adaptors is removed with *Dpn*II. A second set of 12/24 adaptor strands is then added to the amplified tester cDNA population, after which the tester is hybridized against a large excess of driver. The 12mer adaptors are melted and the 3' ends filled in as before. PCR is carried out with primers identical to the new 24mer adaptor. Thus, the only hybridization products which are exponentially amplified are those which are tester:tester combinations. Following PCR, ssDNA products are removed with mung bean nuclease, leaving the 'first difference product'. This is digested and a third set of 12/24 adaptors added before repeating the subtraction process from the hybridization stage. The process is repeated to the 3rd or 4th difference product, as described by Lisitsyn *et al.* (1993) and Hubank and Schatz (1994).

iver and tester cDNA are
et of 12/24 adaptor strands
A products. The 12mer is
A polymerase. Each cDNA
of adaptors is removed with
he amplified tester cDNA
cess of driver. The 12mer
d out with primers identical
which are exponentially

Suppression PCR Subtractive Hybridization (SSH)

The most recent adaptation of the SH approach to differential expression analysis was first described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996). They reported that a 1000–5000 fold enrichment of rare cDNAs (equivalent to isolating mRNAs present at only a few copies per cell) can be obtained without the need for multiple hybridizations/subtractions. Instead of physical or chemical removal of the common sequences, a PCR-based suppression system is used (see figure 5).

In SSH, excess driver cDNA is added to two portions of the tester cDNA which have been ligated with different adaptors. A first round of hybridization serves to enrich differentially expressed genes and equalize rare and abundant messages. Equalization occurs since reannealing is more rapid for abundant molecules than for rarer molecules due to the second order kinetics of hybridization (James and Higgins 1985). The two primary hybridization mixes are then mixed together in the presence of excess driver and allowed to hybridize further. This step permits the annealing of single stranded complementary sequences which did not hybridize in the primary hybridization, and in doing so generates templates for PCR amplification. Although there are several possible combinations of the single stranded molecules present in the secondary hybridization mix, only one particular combination (differentially expressed in the tester cDNA composed of complimentary strands having different adaptors) can amplify exponentially.

Having obtained the final differential display, two options are available if cloning of cDNAs is desired. One is to transform the whole of the final PCR reaction into competent cells. Transformed colonies can then be isolated and their inserts characterized by sequencing, restriction analysis or PCR. Alternatively, the final PCR products can be resolved on a gel and the individual bands excised, reamplified and cloned. The first approach is technically simpler and less time consuming. However, ligation/transformation reactions are known to be biased towards the cloning of smaller molecules, and so the final population of clones will probably not contain a representative selection of the larger products. In addition, although equalization theoretically occurs, observations in this laboratory suggest that this is by no means perfectly accomplished. Consequently, some gene species are present in a higher number than others and this will be represented in the final population of clones. Thus, in order to obtain a substantial proportion of those gene species that actually demonstrate differential expression in the tester population, the number of clones that will have to be screened after this step may be substantial. The second approach is initially more time consuming and technically demanding. However, it would appear to offer better prospects for cloning larger and low abundance gel products. In addition, one can incorporate a screening step that differentiates different products of different sequences but of the same size (HA-staining, see later). In this way, a good idea of the final number of clones to be isolated and identified can be achieved.

An alternative (or even complementary) approach is to use the final differential display reaction to screen a cDNA library to isolate full length clones for further characterization, or a DNA array (see later) to quickly identify known genes. SSH has been used in this laboratory to begin characterization of the short-term gene

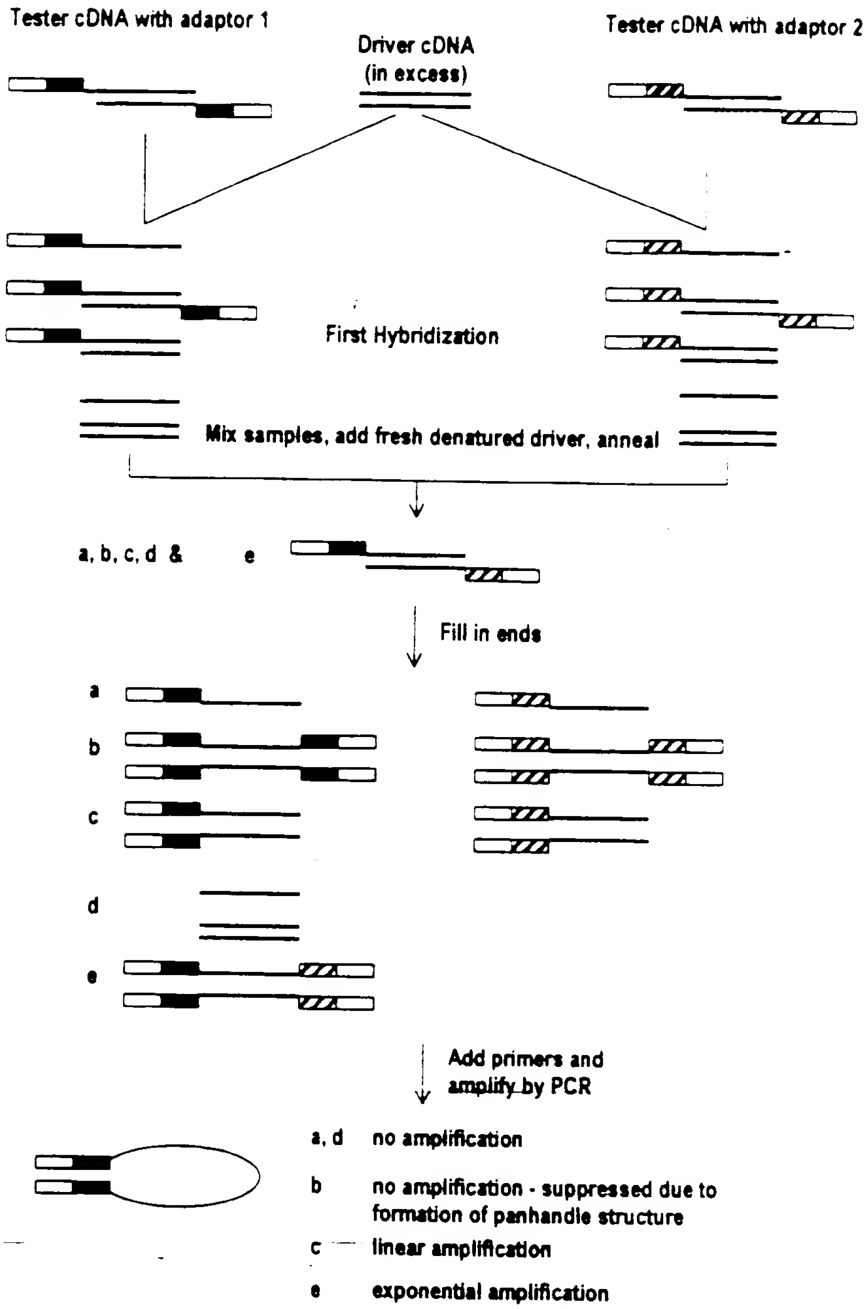

Tester cDNA with adaptor 1
Driver cDNA
(in excess)
Tester cDNA with adaptor 2
First Hybridization
Mix samples, add fresh denatured driver, anneal
a, b, c, d &
e
Fill in ends
a
b
c
d
e
Add primers and
amplify by PCR
a, d
no amplification
b
no amplification - suppressed due to
formation of panhandle structure
c
linear amplification
e
exponential amplification

Figure 5. PCR-select cDNA subtraction. In the primary hybridization, an excess of driver cDNA is added to each tester cDNA population. The samples are heat denatured and allowed to hybridize for between 3 and 8 h. This serves two purposes: (1) to equalize rare and abundant molecules; and (2) to enrich for differentially expressed sequences—cDNAs that are not differentially expressed form type c molecules with the driver. In the secondary hybridization, the two primary hybridizations are mixed together without denaturing. Fresh denatured driver can also be added at this point to allow further enrichment of differentially expressed sequences. Type e molecules are formed in this secondary hybridization which are subsequently amplified using two rounds of PCR. The final products can be visualized on an agarose gel, labelled directly or cloned into a vector for downstream manipulation. As described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996), with permission.

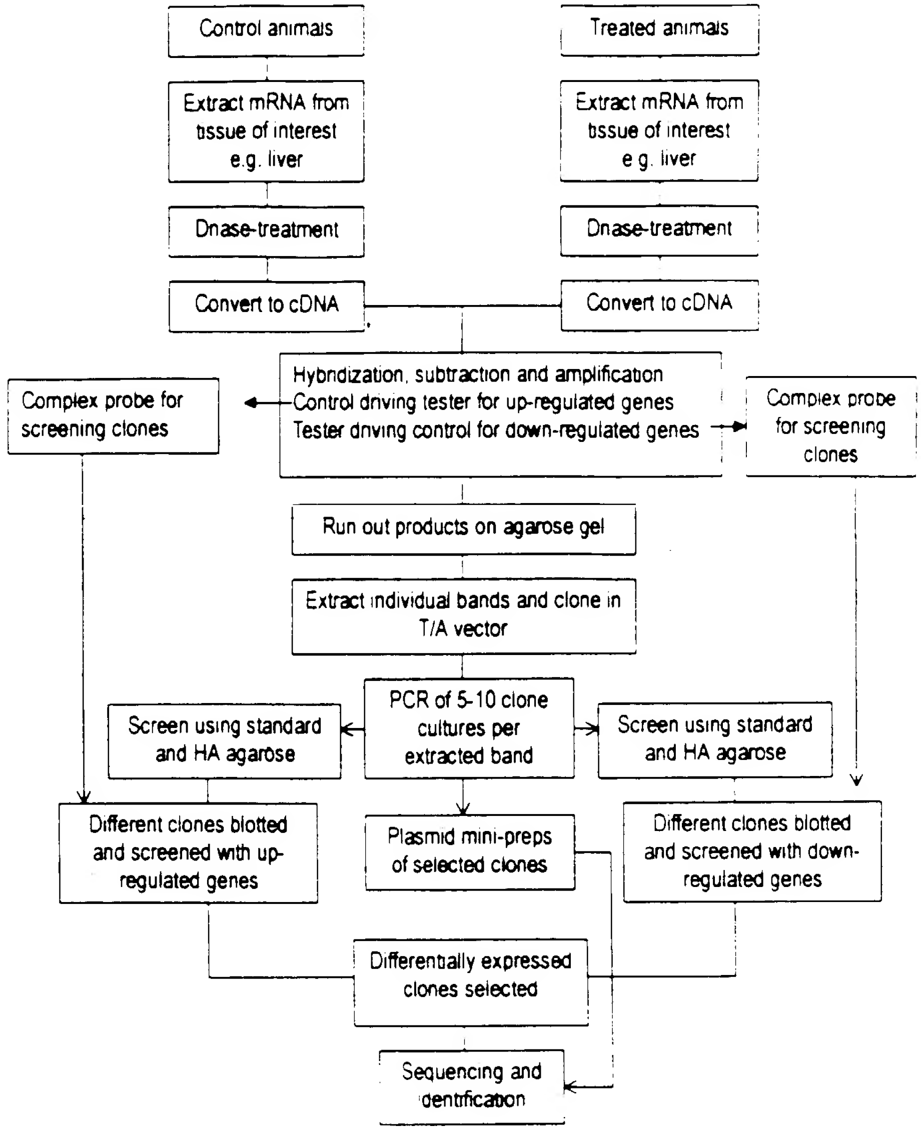

Control animals
Treated animals
Extract mRNA from tissue of interest e.g. liver
Extract mRNA from tissue of interest e.g. liver
Dnase-treatment
Dnase-treatment
Convert to cDNA
Convert to cDNA
Hybridization, subtraction and amplification
Control driving tester for up-regulated genes
Tester driving control for down-regulated genes
Complex probe for screening clones
Complex probe for screening clones
Run out products on agarose gel
Extract individual bands and clone in T/A vector
PCR of 5-10 clone cultures per extracted band
Screen using standard and HA agarose
Screen using standard and HA agarose
Different clones blotted and screened with up-regulated genes
Plasmid mini-preps of selected clones
Different clones blotted and screened with down-regulated genes
Differentially expressed clones selected
Sequencing and identification

Figure 6. Flow diagram showing method used in this laboratory to isolate and identify clones of genes which are differentially expressed in rat liver following short term exposure to the enzyme inducers, phenobarbital and Wy-14,643.

of expressed genes which are unique to each compound and time/dose point. Such information could be useful in short-term characterization of the toxic potential of new compounds by comparing the gene-expression profiles they elicit with those produced by known inducers. Figure 6 shows a flow diagram of the method used to isolate, verify and clone differentially expressed genes, and figure 7 shows expression profiles obtained from a typical SSH experiment. Subsequent sub-cloning of the individual bands, sequencing and gene data base interrogation reveals many genes which are either up- or down-regulated by phenobarbital in the rat (tables 2 and 3).

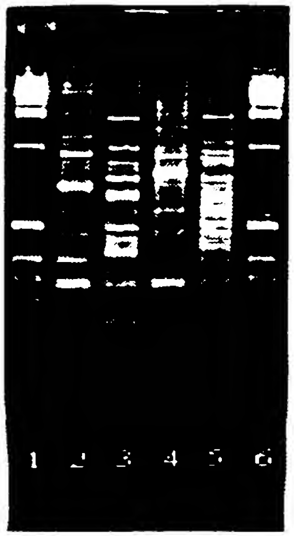


Figure 7. SSH display patterns obtained from rat liver following 3-day treatment with WY-14,643 or phenobarbital. mRNA extracted from control and treated livers was used to generate the differential displays using the PCR-Select cDNA subtraction kit (Clontech). Lane: 1—1kb ladder; 2—genes upregulated following Wy,14-643 treatment; 3—genes downregulated following Wy,14-643 treatment; 4—genes upregulated following phenobarbital treatment; 5—genes downregulated following phenobarbital treatment; 6—1kb ladder. Reproduced from Rockett *et al.* (1997), with permission.

exposure, and an almost complete complement of genes are obtained. For example, the peroxisome proliferator and non-genotoxic hepatocarcinogen Wy,14,643, up-regulates at least 28 genes and down-regulates at least 15 in the rat (a sensitive species) and produces 48 up- and 37 down-regulated genes in the guinea pig, a resistant species (Rockett, Swales, Esda and Gibson, unpublished observations). One of these genes, CD81, was up-regulated in the rat and down-regulated in the guinea pig following Wy-14,643 treatment. CD81 (alternatively named TAPA-1) is a widely expressed cell surface protein which is involved in a large number of cellular processes including adhesion, activation, proliferation and differentiation (Levy *et al.* 1998). Since all of these functions are altered to some extent in the phenomena of hepatomegaly and non-genotoxic hepatocarcinogenesis, it is intriguing, and probably mechanistically-relevant, that CD81 expression is differentially regulated in a resistant and susceptible species. However, the down-side of this approach is that the majority of genes can be sequenced and matched to database sequences, but the latter are predominantly expressed sequence tags or genes of completely unknown function, thus partially obscuring a realistic overall assessment of the critical genes of genuine biological interest. Notwithstanding the lack of complete funtional identification of altered gene expression, such gene profiling studies essentially provides a 'molecular fingerprint' in response to xenobiotic challenge, thereby serving as a mechanistically-relevant platform for further detailed investigations.

Differential Display (DD)

Originally described as 'RNA fingerprinting by arbitrarily primed PCR' (Liang and Pardee 1992) this method is now more commonly referred to as 'differential

atment with WY-14,643 or
was used to generate the
(Clontech). Lane: 1—1kb
es downregulated following
rbital treatment; 5—genes
Reproduced from Rockett *et*

btained. For example,
nogen Wy,14,643, up-
in the rat (a sensitive
s in the guinea pig, a
blished observations).
down-regulated in the
ely named TAPA-1) is
arge number of cellular
ifferentiation (Levy *et*
ent in the phenomena
it is intriguing, and
ifferentially regulated
de of this approach is
atabase sequences, but
genes of completely
rall assessment of the
g the lack of complete
gene profiling studies
xenobiotic challenge,
for further detailed

Table 2. Genes up-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity		FASTA-EMBL gene identification
5 (1300)		93.5%	CYP2B1
7 (1000)		95.1%	Preproalbumin Serum albumin mRNA
8 (950)		98.3%	NCI-CGAP-Pr1 *H. sapiens* (EST)
10 (850)		95.7%	CYP2B1
11 (800)	Clone 1	94.9%	CYP2B1
	Clone 2	75.3%	CYP2B2
12 (750)		93.8%	TRPM-2 mRNA Sulfated glycoprotein
15 (600)		92.9%	Preproalbumin Serum albumin mRNA
16 (55)	Clone 1	95.2%	CYP2B1
	Clone 2	93.6%	Haptoglobulin mRNA partial alpha
21 (350)		99.3%	18S, 5.8S & 28S rRNa

Bands 1–4, 6, 9, 13, 14, and 17–20 are shown to be false positives by dot blot analysis and, therefore, are not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are up-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

Table 3. Genes down-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity		FASTA-EMBL gene identification
1 (1500)		95.3%	3-oxoacyl-CoA thiolase
2 (1200)		92.3%	Hemopoxin mRNA
3 (1000)		91.7%	Alpha-2u-globulin mRNA
7 (700)	Clone 1	77.2%	*M.musculus* C1 inhibitor
	Clone 2	94.5%	Electron transfer flavoprotein
	Clone 3	91.0%	*M. musculus* Topoisomerase 1 (Topo 1)
8 (650)	Clone 1	86.9%	Soares 2NbMT *M. musculus* (EST)
	Clone 2	96.2%	Alpha-2u-globulin (s-type) mRNA
9 (600)	Clone 1	86.9%	Soares mouse NML *M. musculus* (EST)
	Clone 2	82.0%	Soares p3NMF 19.5 *M. musculus* (EST)
10 (550)		73.8%	Soares mouse NML *M. musculus* (EST)
11 (525)		95.7%	NCI-CGAP-Pr1 *H. sapiens* (EST)
12 (375)		100.0%	Ribosomal protein
13 (23)	Clone 1	97.2%	Soares mouse embryo NbME135 (EST)
	Clone 2	100.0	Fibrinogen B-beta-chain
	Clone 3	100.0%	Apolipoprotein E gene
14 (170)		96.0%	Soares p3NMF19.5 *M. musculus* (EST)
15 (140)		97.3%	Stratagene mouse testis (EST)
Others: (300)		96.7%	*R. norvegicus* RASP 1 mRNA
(275)		93.1%	Soares mouse mammary gland (EST)

EST = Expressed sequence tag. Bands 4–6 were shown to be false positives by dot blot analysis and, therefore, were not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are down-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

display' (DD). In this method, all the mRNA species in the control and treated cell populations are amplified in separate reactions using reverse transcriptase-PCR display compared to the other, are differentially expressed and may be recovered for further characterization. One advantage of this system is the speed with which it can be carried out—2 days to obtain a display and as little as a week to make and identify clones.

Two commonly used variations are based on different methods of priming the reverse transcription step (figure 8). One is to use an oligo dT with a 2-base 'anchor' at the 3'-end, e.g. 5' $(dT_{11})CA$ 3' (Liang and Pardee 1992). Alternatively, an arbitrary primer may be used for 1st strand cDNA synthesis (Welsh *et al.* 1992). This variant of RNA fingerprinting has also been called 'RAP' (RNA Arbitrarily Primed)-PCR. One advantage of this second approach is that PCR products may be derived from anywhere in the RNA, including open reading frames. In addition, it can be used for mRNAs that are not polyadenylated, such as many bacterial mRNAs (Wong and McClelland 1994). In both cases, following reverse transcription and denaturation, second strand cDNA synthesis is carried out with an arbitrary primer (*arbitrary* primers have a single base at each position, as compared to *random* primers, which contain a mixture of all four bases at each position). The resulting PCR, thus, produces a series of products which, depending on the system (primer length and composition, polymerase and gel system), usually includes 50–100 products per primer set (Band and Sager 1989). When a combination of different dT-anchors and arbitrary primers are used, almost all mRNA species from a cell can be amplified. When the cDNA products from two different populations are analysed side by side on a polyacrylamide gel, differences in expression can be identified and the appropriate bands recovered for cloning and further analysis.

Although DD is perhaps the most popular approach used today for identifying differentially expressed genes, it does suffer from several perceived disadvantages:

(1) It may have a strong bias towards high copy number mRNAs (Bertioli *et al.* 1995), although this has been disputed (Wan *et al.* 1996) and the isolation of very low abundance genes may be achieved in certain circumstances (Guimeraes *et al.* 1995a).

(2) The cDNAs obtained often only represent the extreme 3' end of the mRNA (often the 3'-untranslated region), although this may not always be the case (Guimeraes *et al.* 1995a). Since the 3' end is often not included in Genbank and shows variation between organisms, cDNAs identified by DD cannot always be matched with their genes, even if they have been identified.

(3) The pattern of differential expression seen on the display often cannot be reproduced on Northern blots, with false positives arising in up to 70 % of cases (Sun *et al.* 1994). Some adaptations have been shown to reduce false positives, including the use of two reverse transcriptases (Sung and Denman 1997), comparison of uninduced and induced cells over a time course (Burn *et al.* 1994) and comparison of DDPCR-products from two uninduced and two induced lines (Sompayrac *et al.* 1995). The latter authors also reported that the use of cytoplasmic RNA rather then total RNA reduces false positives arising from nuclear RNA that is not transported to the cytoplasm.

Further details of the background, strengths and weaknesses of the DD technique can be obtained from a review by McClelland *et al.* (1996) and from articles by Liang *et al.* (1995) and Wan *et al.* (1996).

d may be recovered for
peed with which it can
k to make and identify

ethods of priming the
with a 2-base 'anchor'
92). Alternatively, an
is (Welsh *et al.* 1992).
AP' (RNA Arbitrarily
PCR products may be
frames. In addition, it
iany bacterial mRNAs
erse transcription and
ith an arbitrary primer
compared to *random*
osition). The resulting
on the system (primer
ially includes 50–100
mbination of different
species from a cell can
pulations are analysed
n can be identified and
lysis.
l today for identifying
ceived disadvantages:

iRNAs (Bertioli *et al.*
id the isolation of very
stances (Guimeraes *et*

3′ end of the mRNA
ot always be the case
uded in Genbank and
DD cannot always be
ed.
play often cannot be
in up to 70% of cases
reduce false positives,
and Denman 1997),
urse (Burn *et al.* 1994)
ced and two induced
ported that the use of
positives arising from

aknesses of the DD

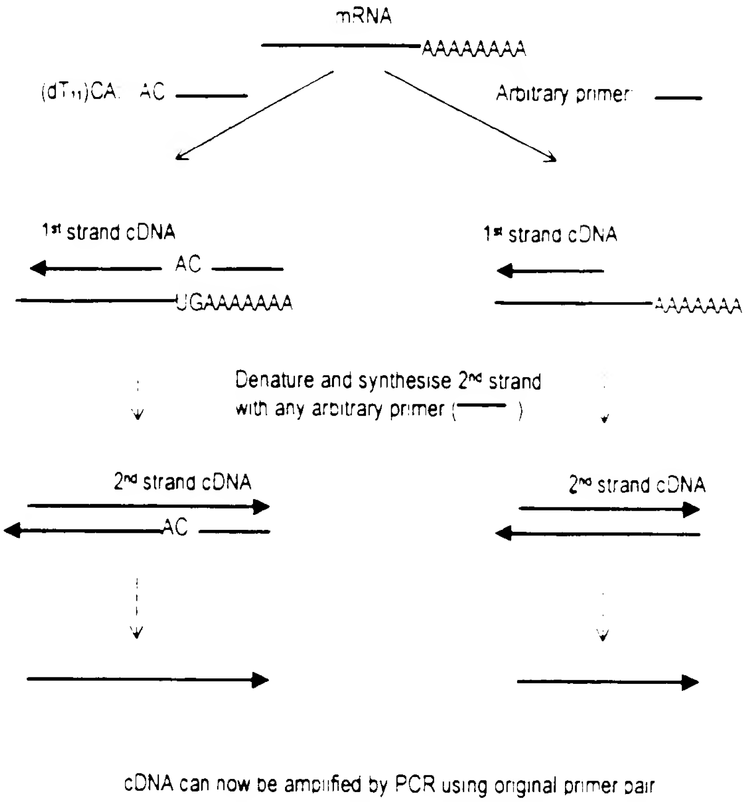

mRNA
AAAAAAAA
(dT11)CA AC
Arbitrary primer
1st strand cDNA
AC
UGAAAAAAA
1st strand cDNA
AAAAAAA
Denature and synthesise 2nd strand
with any arbitrary primer ( )
2nd strand cDNA
AC
2nd strand cDNA
cDNA can now be amplified by PCR using original primer pair

Figure 8. Two approaches to differential display (DD) analysis. 1st strand synthesis can be carried out either with a polydT$_{11}$NN primer (where N = G, C or A) or with an arbitrary primer. The use of different combinations of G, C and A to anchor the first strand polydT primer enables the priming of the majority of polyadenylated mRNAs. Arbitrary primers may hybridize at none, one or more places along the length of the mRNA, allowing 1st strand cDNA synthesis to occur at none, one or more points in the same gene. In both cases, 2nd strand synthesis is carried out with an arbitrary primer. Since these arbitrary primers for the 2nd strand may also hybridize to the 1st strand cDNA in a number of different places, several different 2nd strand products may be obtained from one binding point of the 1st strand primer. Following 2nd strand synthesis, the original set of primers is used to amplify the second strand products, with the result that numerous gene sequences are amplified.

Restriction endonuclease-facilitated analysis of gene expression

Serial Analysis of Gene Expression (SAGE)

A more recent development in the field of differential display is SAGE analysis (Velculescu *et al.* 1995). This method uses a different approach to those discussed so far and is based on two principles. Firstly, in more than 95% of cases, short nucleotide sequences ('tags') of only nine or 10 base pairs provide sufficient information to identify their gene of origin. Secondly, concatonation (linking together in a series) of these tags allows sequencing of multiple cDNAs within a single clone. Figure 9 shows a schematic representation of the SAGE process. In this

cDNA population are captured with streptavidin beads. The captured population split into two and different adaptors ligated to the 5′ ends of each group. Incorporated into the adaptors is a recognition sequence for a type IIS restriction enzyme—one which cuts DNA at a defined distance (< 20 bp) from its recognition sequence. Hence, following digestion of each captured cDNA population with the IIS enzyme, the adaptors plus a short piece of the captured cDNA are released. The two populations are then ligated and the products amplified. The amplified products are cleaved with the original anchoring enzyme, religated (concatomers are formed in the process) and cloned. The advantage of this system is that hundreds of gene tags can be identified by sequencing only a few clones. Furthermore, the number of times a given transcript is identified is a quantitative measurement of that gene's abundance in the original population, a feature which facilitates identification of differentially expressed genes in different cell populations.

Some disadvantages of SAGE analysis include the technical difficulty of the method, a large amount of accurate sequencing is required, biased towards abundant mRNAs, has not been validated in the pharmaco/toxicogenomic setting and has only been used to examine well known tissue differences to date.

Gene Expression Fingerprinting (GEF)

A different capture/restriction digest approach for isolating differentially expressed genes has been described by Ivanova and Belyavsky (1995). In this method, RNA is converted to cDNA using biotinylated oligo(dT) primers. The cDNA population is then digested with a specific endonuclease and captured with magnetic streptavidin microbeads to facilitate removal of the unwanted 5′ digestion products. The use of restricted 3′-ends alone serves to reduce the complexity of the cDNA fragment pool and helps to ensure that each RNA species is represented by not more than one restriction product. An adaptor is ligated to facilitate subsequent amplification of the captured population. PCR is carried out with one adaptor-specific and one biotinylated polydT primer. The reamplified population is recaptured and the non-biotinylated strands removed by alkaline dissociation. The non-biotinylated strand is then resynthesized using a different adaptor-specific primer in the presence of a radiolabelled dNTP. The labelled immobilized 3′ cDNA ends are next sequentially treated with a series of different restriction endonucleases and the products from each digestion analysed by PAGE. The result is a fingerprint composed of a number of ladders (equal to the number of sequential digests used). By comparing test versus control fingerprints, it is possible to identify differentially expressed products which can then be isolated from the gel and cloned. The advantages of this procedure are that it is very robust and reproducible, and the authors estimate that 80–93% of cDNA molecules are involved in the final fingerprint. The disadvantage is that polyacrylamide gels can rarely resolve more than 300–400 bands, which compares poorly to the 1000 or more which are estimated to be produced in an average experiment. The use of 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991) may help to overcome this problem.

A similar method for displaying restriction endonuclease fragments was later described by Prashar and Weissman (1996). However, instead of sequential digestion of the immobolized 3'-terminal cDNA fragments, these authors simply compared the profiles of the control and treated populations without further manipulation.

ch group. Incorporated
striction enzyme—one
recognition sequence.
n with the IIS enzyme,
ire released. The two
amplified products are
atomers are formed in
: hundreds of gene tags
re, the number of times
ement of that gene's
litates identification of

hnical difficulty of the
ased towards abundant
nomic setting and has
date.

isolating differentially
avsky (1995). In this
ligo(dT) primers. The
ease and captured with
unwanted 5′ digestion
e the complexity of the
ecies is represented by
to facilitate subsequent
out with one adaptor-
nplified population is
aline dissociation. The
ferent adaptor-specific
immobilized 3′ cDNA
striction endonucleases
e result is a fingerprint
quential digests used).
) identify differentially
gel and cloned. The
reproducible, and the
involved in the final
an rarely resolve more
0 or more which are
se of 2-D gels such as
al. (1991) may help to

se fragments was later
instead of sequential

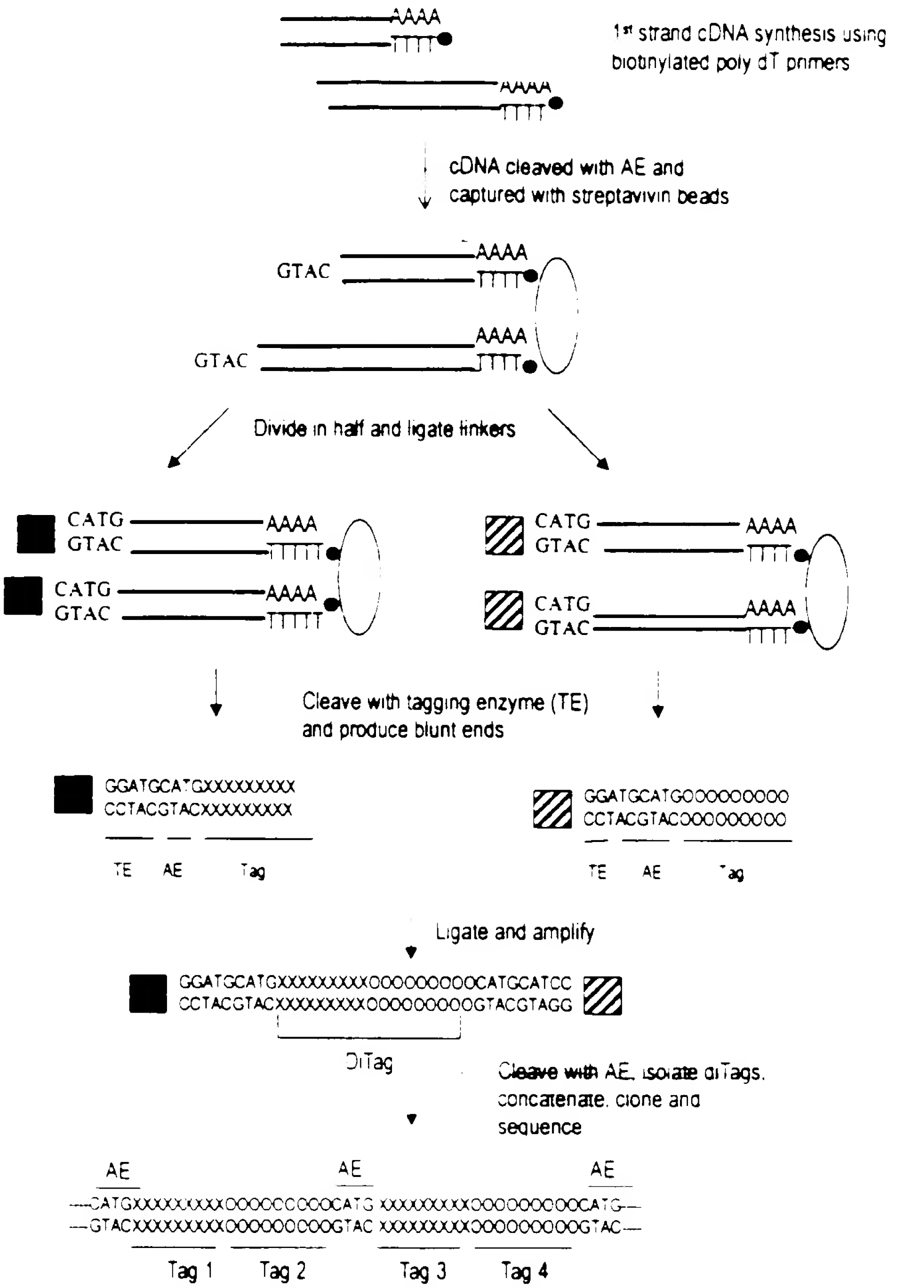

AAAA
TTTT
1st strand cDNA synthesis using biotinylated poly dT primers
cDNA cleaved with AE and captured with streptavivin beads
GTAC
AAAA
TTTT
Divide in half and ligate linkers
CATG
GTAC
Cleave with tagging enzyme (TE) and produce blunt ends
GGATGCATGXXXXXXXXX
CCTACGTACXXXXXXXXX
TE
AE
Tag
GGATGCATGOOOOOOOOO
CCTACGTACOOOOOOOOO
Ligate and amplify
GGATGCATGXXXXXXXXXOOOOOOOOOCATGCATCC
CCTACGTACXXXXXXXXXOOOOOOOOOGTACGTAGG
DiTag
Cleave with AE, isolate diTags, concatenate, clone and sequence
AE
—CATGXXXXXXXXXOOOOOOOOOCATG XXXXXXXXXOOOOOOOOOCATG—
—GTACXXXXXXXXXOOOOOOOOOGTAC XXXXXXXXXOOOOOOOOOGTAC—
Tag 1
Tag 2
Tag 3
Tag 4

Figure 9. Serial analysis of gene expression (SAGE) analysis. cDNA is cleaved with an anchoring enzyme (AE) and the 3′ends captured using streptavidin beads. The cDNA pool is divided in half and each portion ligated to a different linker, each containing a type IIS restriction site (tagging enzyme, TE). Restriction with the type IIS enzyme releases the linker plus a short length of cDNA (XXXXX and OOOOO indicate nucleotides of different tags). The two pools of tags are then ligated and amplified using linker-specific primers. Following PCR, the products are cleaved with the AE and the ditags isolated from the linkers using PAGE. The ditags are then ligated (during

DNA arrays

'Open' differential display systems are cumbersome in that it takes a great deal of time to extract and identify candidate genes and then confirm that they are indeed up- or down-regulated in the treated compared to the control tissue. Normally, the latter process is carried out using Northern blotting or RT-PCR. Even so, each of the aforementioned steps produce a bottleneck to the ultimate goal of rapid analysis of gene expression. These problems will likely be addressed by the development of so-called DNA arrays (e.g. Gress *et al.* 1992, Zhao *et al.* 1995, Schena *et al.* 1996), the introduction of which has signalled the next era in differential gene expression analysis. DNA arrays consist of a gridded membrane or glass 'chips' containing hundreds or thousands of DNA spots, each consisting of multiple copies of part of a known gene. The genes are often selected based on previously proven involvement in oncogenesis, cell cycling, DNA repair, development and other cellular processes. They are usually chosen to be as specific as possible for each gene and animal species. Human and mouse arrays are already commercially available and a few companies will construct a personalized array to order, for example Clontech Laboratories and Research Genetics Inc. The technique is rapid in that hundreds or even thousands of genes can be spotted on a single array, and that mRNA/cDNA from the test populations can be labelled and used directly as probe. When analysed with appropriate hardware and software, arrays offer a rapid and quantitative means to assess differences in gene expression between two cell populations. Of course, there can only be identification and quantitation of those genes which are in the array (hence the term 'closed' system). Therefore, one approach to elucidating the molecular mechanisms involved in a particular disease/development system may be to combine an open and closed system—a DNA array to directly identify and quantitate the expression of known genes in mRNA populations, and an open system such as SSH to isolate unknown genes which are differentially expressed.

One of the main advantages of DNA arrays is the huge number of gene fragments which can be put on a membrane—some companies have reported gridding up to 60000 spots on a single glass 'chip' (microscope slide). These high density chip-based micro-arrays will probably become available as mass-produced off-the-shelf items in the near future. This should facilitate the more rapid determination of differential expression in time and dose-response experiments. Aside from their high cost and the technical complexities involved in producing and probing DNA arrays, the main problem which remains, especially with the newer micro-array (gene-chip) technologies, is that results are often not wholly reproducible between arrays. However, this problem is being addressed and should be resolved within the next few years.

EST databases as a means to identify differentially expressed genes

Expressed sequence tags (ESTs) are partial sequences of clones obtained from cDNA libraries. Even though most ESTs have no formal identity (putative identification is the best to be hoped for), they have proven to be a rapid and efficient means of discovering new genes and can be used to generate profiles of gene-expression in specific cells. Since they were first described by Adams *et al.* (1991), there has been a huge explosion in EST production and it is estimated that there are now well over a million such sequences in the public domain, representing over half

hat it takes a great deal
rm that they are indeed
l tissue. Normally, the
PCR. Even so, each of
e goal of rapid analysis
by the development of
'5, Schena *et al.* 1996),
ential gene expression
ass 'chips' containing
ultiple copies of part of
ly proven involvement
ther cellular processes.
ene and animal species.
e and a few companies
itech Laboratories and
eds or even thousands
./cDNA from the test
When analysed with
quantitative means to
tions. Of course, there
which are in the array
ch to elucidating the
opment system may be
directly identify and
ulations, and an open
ferentially expressed.
nber of gene fragments
ported gridding up to
ese high density chip-
produced off-the-shelf
apid determination of
nts. Aside from their
ng and probing DNA
he newer micro-array
reproducible between
be resolved within the

ressed genes

clones obtained from
nal identity (putative
be a rapid and efficient
rate profiles of gene-
[illegible] 1991

of all human genes (Hillier *et al.* 1996). This large number of freely available sequences (both sequence information and clones are normally available royalty-free from the originators) has enabled the development of a new approach towards differential gene expression analysis as described by Vasmatzis *et al.* (1998). The approach is simple in theory: EST databases are first searched for genes that have a number of related EST sequences from the target tissue of choice, but none or few from non-target tissue libraries. Programmes to assist in the assembly of such sets of overlapping data may be developed in-house or obtained privately or from the internet. For example, the Institute for Genomic Research (TIGR, found at http://www.tigr.org) provides many software tools free of charge to the scientific community. Included amongst these is the TIGR assembler (Sutton *et al.* 1995), a tool for the assembly of large sets of overlapping data such as ESTs, bacterial artificial chromosomes (BAC)s, or small genomes. Candidate EST clones representing different genes are then analysed using RNA blot methods for size and tissue specificity and, if required, used as probes to isolate and identify the full length cDNA clone for further characterization. In practice however, the method is rather more involved, requiring bioinformatic and computer analysis coupled with confirmatory molecular studies. Vasmatzis *et al.* (1998) have described several problems in this fledgling approach, such as separating highly homologous sequences derived from different genes and an overemphasis of specificity for some EST sequences. However, since these problems will largely be addressed by the development of more suitable computer algorithms and an increased completeness of the EST database, it is likely that this approach to identifying differentially expressed genes may enjoy more patronage in the future.

Problems and potential of differential expression techniques

The holistic or single cell approach?

When working with *in vivo* models of differential expression, one of the first issues to consider must be the presence of multiple cell types in any given specimen. For example, a liver sample is likely to contain not only hepatocytes, but also (potentially) Ito cells, bile ductule cells, endothelial cells, various immune cells (e.g. lymphocytes, macrophages and Kupffer cells) and fibroblasts. Other tissues will each have their own distinctive cell populations. Also, in the case of neoplastic tissue, there are almost always normal, hyperplastic and or dysplastic cells present in a sample. One must, therefore, be aware that genes obtained from a differential display experiment performed on an animal tissue model may not necessarily arise exclusively from the intended 'target' cells, e.g. hepatocytes neoplastic cells. If appropriate, further analyses using immunohistochemistry, *in situ* hybridization or ***in situ*** RT-PCR should be used to confirm which cell types are expressing the gene(s) of interest. This problem is probably most acute for those studying the differential expression of genes in the development of different cell types, where there is a need to examine homologous cell populations. The problem is now being addressed at the National Cancer Institute (Bethesda, MD, USA) where new micro-disection techniques have been employed to assist in their gene analysis programme, the Cancer Genome Anatomy Project (CGAP) (For more information see web site e.g. fluorescence activated cell sorting (FACS) (Dunbar *et al.* 1998, Kas-Deelen *et al.* 1998) and magnetic bead technology (Richard *et al.* 1998, Rogler *et al.* 1998).

However, those taking a holistic approach may consider this issue unimportant. There is an equally appropriate view that all those genes showing altered expression within a compromized tissue should be taken into consideration. After all, since all tissues are complex mixes of different, interacting cell types which intimately regulate each other's growth and development, it is clear that each cell type could in some way contribute (positively or negatively) towards the molecular mechanisms which lie behind responses to external stimuli or neoplastic growth It is perhaps then more informative to carry out differential display experiments using *in vivo* as opposed to *in vitro* models, where uniform populations of identical cells probably represent a partial, skewed or even inaccurate picture of the molecular changes that occur.

The incidence and possible implications of inter-individual biological variation should be considered in any approach where whole animal models are being used. It is clear that individuals (humans and animals) respond in different ways to identical stimuli. One of the best characterized examples is the debrisoquine oxidation polymorphism, which is mediated by cytochrome CYP2D6 and determines the pharmacokinetics of many commonly prescribed drugs (Lennard 1993, Meyer and Zanger 1997). The reasons for such differences are varied and complex, but allelic variations, regulatory region polymorphisms and even physical and mental health can all contribute to observed differences in individual responses. Careful thought should, therefore, be given to the specific objectives of the study and to the possible value of pooling starting material (tissue/mRNA). The effect of this can be beneficial through the ironing out of exaggerated responses and unimportant minor fluctuations of (mechanistically) irrelevant genes in individual animals, thus providing a clearer overall picture of the general molecular mechanisms of the response. However, at the same time such minor variations may be of utmost importance in deciding the ability of individual animals to succumb to or resist the effects of a given chemical/disease.

How efficient are differential expression techniques at recovering a high percentage of differentially expressed genes?

A number of groups have produced experimental data suggesting that mammalian cells produce between 8000–15000 different mRNA species at any one time (Mechler and Rabbitts 1981, Hedrick *et al.* 1984, Bravo 1990), although figures as high as 20–30000 have also been quoted (Axel *et al.* 1976). Hedrick *et al.* (1984) provided evidence suggesting that the majority of these belong to the rare abundance class. A breakdown of this abundance distribution is shown in table 1.

When the results of differential display experiments have been compared with data obtained previously using other methods, it is apparent that not all differentially expressed mRNAs are represented in the final display. In particular, rare messages (which, importantly, often include regulatory proteins) are not easily recovered using differential display systems. This is a major shortcoming, as the majority of mRNA species exist at levels of less than 0.005% of the total population (table 1). Bertioli *et al.* (1995) examined the efficiency of DD templates (heterogeneous mRNA populations) for recovering rare messages and were unable to detect mRNA

. 1998, Kas-Deelen *et*
, Rogler *et al.* 1998).
his issue unimportant.
ing altered expression
ion. After all, since all
pes which intimately
each cell type could in
nolecular mechanisms
growth. It is perhaps
ments using *in vivo* as
lentical cells probably
nolecular changes that

ial biological variation
dels are being used. It
erent ways to identical
ebrisoquine oxidation
6 and determines the
nard 1993, Meyer and
d complex, but allelic
cal and mental health
nses. Careful thought
dy and to the possible
effect of this can be
d unimportant minor
vidual animals, thus
r mechanisms of the
ns may be of utmost
ccumb to or resist the

a high percentage of

uggesting that mam-
pecies at any one time
), although figures as
Hedrick *et al.* (1984)
to the rare abundance
in table 1.
been compared with
at not all differentially
ticular, rare messages
not easily recovered
ng, as the majority of

species present at less than 1.2% of the total mRNA population—equivalent to an intermediate or abundant species. Interestingly, when simple model systems (single target only) were used instead of a heterogeneous mRNA population, the same primers could detect levels of target mRNA down to 10000 × smaller. These results are probably best explained by competition for substrates from the many PCR products produced in a DD reaction.

The numbers of differentially expressed mRNAs reported in the literature using various model systems provides further evidence that many differentially expressed mRNAs are not recovered. For example, DeRisi *et al.* (1997) used DNA array technology to examine gene expression in yeast following exhaustion of sugar in the medium, and found that more than 1700 genes showed a change in expression of at least 2-fold. In light of such a finding, it would not be unreasonable to suggest that of the 8000–15 000 different mRNA species produced by any given mammalian cell, up to 1000 or more may show altered expression following chemical stimulation. Whilst this may be an extreme figure, it is known that at least 100 genes are activated/upregulated in Jurkat (T-) cells following IL-2 stimulation (Ullman *et al.* 1990). In addition, Wan *et al.* (1996) estimated that interferon-γ-stimulated HeLa cells differentially express up to 433 genes (assuming 24000 distinct mRNAs expressed by the cells). However, there have been few publications documenting anywhere near the recovery of these numbers. For example, in using DD to compare normal and regenerating mouse liver, Bauer *et al.* (1993) found only 70 of 38000 total bands to be different. Of these, 50% (35 genes) were shown to correspond to differentially expressed bands. Chen *et al.* (1996) reported 10 genes upregulated in female rat liver following ethinyl estradiol treatment. McKenzie and Drake (1997) identified 14 different gene products whose expression was altered by phorbol myristate acetate (PMA, a tumour promoter agent) stimulation of a human myelomonocytic cell line. Kilty and Vickers (1997) identified 10 different gene products whose expression was upregulated in the peripheral blood leukocytes of allergic disease sufferers. Linskens *et al.* (1995) found 23 genes differentially expressed between young and senescent fibroblasts. Techniques other than DD have also provided an apparent paucity of differentially expressed genes. Using SH for example, Cao *et al.* (1997) found 15 genes differentially expressed in colorectal cancer compared to normal mucosal epithelium. Fitzpatrick *et al.* (1995) isolated 17 genes upregulated in rat liver following treatment with the peroxisome proliferator, clofibrate; Philips *et al.* (1990) isolated 12 cDNA clones which were upregulated in highly metastatic mammary adenocarcinoma cell lines compared to poorly metastatic ones. Prashar and Weissman (1996) used 3′ restriction fragment analysis and identified approximately 40 genes showing altered expression within 4 h of activation of Jurkat T-cells. Groenink and Leegwater (1996) analysed 27 gene fragments isolated using SSH of delayed early response phase of liver regeneration and found only 12 to be upregulated.

In the laboratory, SSH was used to isolate up to 70 candidate genes which appear to show altered expression in guinea pig liver following short-term treatment with the peroxisome proliferator, WY-14,643 (Rockett, Swales, Esdaile and Gibson, unpublished observations). However, these findings have still to be confirmed by analysis of the extracted tissue mRNA for differential expression of these sequences.

positives), it is still not clear if such adaptations are practically effective—proving efficiency by spiking with a known amount of limited numbers of artificial construct(s) is one thing, but isolating a high percentage of the rare messages already present in an mRNA population is another. Of course, some models will genuinely produce only a small number of differentially expressed genes. In addition, there are also technical problems that can reduce efficiency. For example, mRNAs may have an unusual primary structure that effectively prevents their amplification by PCR-based systems. In addition, it is known that under certain circumstances not all mRNAs have 3′ polyA sites. For example, during *Xenopus* development, deadenylation is used as a means to stabilize RNAs (Voeltz and Steitz 1998), whilst preferential deadenylation may play a role in regulating Hsp70 (and perhaps, therefore, other stress protein) expression in *Drosophila* (Dellavalle *et al.* 1994). The presence of deadenylated mRNAs would clearly reduce the efficiency of systems utilizing a polydT reverse transcription step. The efficiency of any system also depends on the quality of the starting material. All differential display techniques use mRNA as their target material. However, it is difficult to isolate mRNA that is completely free of ribosomal RNA. Even if polydT primers are used to prime first strand cDNA synthesis, ribosomal RNA is often transcribed to some degree (Clontech PCR-Select cDNA Subtraction kit user manual). It has been shown, at least in the case of SSH, that a high rRNA:mRNA ratio can lead to inefficient subtractive hybridization (Clontech PCR-Select cDNA Subtraction kit user manual), and there is no reason to suppose that it will not do likewise in other SH approaches. Finally, those techniques that utilise a presubtraction amplification step (e.g. RDA) may present a skewed representation since some sequences amplify better than others.

Of course, probably the most important consideration is the temporal factor. It is clear that any given differential display experiment can only interrogate a cell at one point in time. It may well be that a high percentage of the genes showing altered expression at that time are obtained. However, given that disease processes and responses to environmental stimuli involve dynamic cascades of signalling, regulation, production and action, it is clear that all those genes which are switched on/off at different times will not be recovered and, therefore, vital information may well be missed. It is, therefore, imperative to obtain as much information about the model system beforehand as possible, from which a strategy can be derived for targeting specific time points or events that are of particular interest to the investigator. One way of getting round this problem of single time point analysis is to conduct the experiment over a suitable time course which, of course, adds substantially to the amount of work involved.

How sensitive are differential expression technologies?

There has been little published data that addresses the issue of how large the change in expression must be for it to permit isolation of the gene in question with the various differential expression technologies. Although the isolation of genes whose expression is changed as little as 1.5-fold has been reported using SSH (Groenink and Leegwater 1996), it appears that those demonstrating a change in excess of 5-fold are more likely to be picked up. Thus, there is a 'grey zone' in between where small changes could fade in and out of isolation between

ally effective—proving
numbers of artificial
e rare messages already
: models will genuinely
s. In addition, there are
ple, mRNAs may have
amplification by PCR-
circumstances not all
evelopment, deadenyl-
d Steitz 1998), whilst
Hsp70 (and perhaps,
lavalle *et al.* 1994). The
e efficiency of systems
cy of any system also
tial display techniques
o isolate mRNA that is
are used to prime first
ribed to some degree
It has been shown, at
can lead to inefficient
Subtraction kit user
o likewise in other SH
ction amplification step
me sequences amplify

the temporal factor. It
ıly interrogate a cell at
genes showing altered
disease processes and
ascades of signalling,
nes which are switched
vital information may
information about the
gy can be derived for
cular interest to the
e time point analysis is
hich, of course, adds

issue of how large the
gene in question with
the isolation of genes
reported using SSH

experiments and animals. DD, on the other hand, is not subject to this grey zone since, unlike SH approaches, it does not amplify the difference in expression between two samples. Wan *et al.* (1996) reported that differences in expression of twofold or more are detectable using DD.

Resolution and visualization of differential expression products

It seems highly improbable with current technology that a gel system could be developed that is able to resolve all gene species showing altered expression in any given test system (be it SH- or DD-based). Polyacrylamide gel electrophoresis (PAGE) can resolve size differences down to 0.2% (Sambrook *et al.* 1989) and are used as standard in DD experiments. Even so, it is clear that a complex series of gene products such as those seen in a DD will contain unresolvable components. Thus, what appears to be one band in a gel may in fact turn out to be several. Indeed, it has been well documented (Mathieu-Daude *et al.* 1996, Smith *et al.* 1997) that a single band extracted from a DD often represents a composite of heterogeneous products, and the same has been found for SSH displays in this laboratory (Rockett *et al.* 1997). One possible solution was offered by Mathieu-Daude *et al.* (1996), who extracted and reamplified candidate bands from a DD display and used single strand conformation polymorphism (SSCP) analysis to confirm which components represented the truly differentially expressed product.

Many scientists often try to avoid the use of PAGE where possible because it is technically more demanding than agarose gel electrophoresis (AGE). Unfortunately, high resolution agarose gels such as Metaphor (FMC, Lichfield, UK) and AquaPor HR (National Diagnostics, Hessle, UK), whilst easier to prepare and manipulate than PAGE, can only separate DNA sequences which differ in size by around 1.5–2% (15–20 base pairs for a 1Kb fragment). Thus, SSH, RDA or other such products which differ in size by less than this amount are normally not resolvable. However, a simple technique does in fact exist for increasing the resolving power of AGE—the inclusion of HA-red (10-phenyl neutral red-PEG ligand) or HA-yellow (bisbenzamide-PEG ligand) (Hanse Analytik GmbH, Bremen, Germany) in a gel separates identical or closely sized products on base content. Specifically, HA-red and -yellow selectively bind to GC and AT DNA motifs, respectively (Wawer *et al.* 1995, Hanse Analytik 1997, personal communication). Since both HA-stains possess an overall positive charge, they migrate towards the cathode when an electric field is applied. This is in direct opposition to DNA, which is negatively charged and, therefore, migrates towards the anode. Thus, if two DNA clones are identical in size (as perceived on a standard high resolution agarose gel), but differ in AT/GC content, inclusion of a HA-dye in the gel will effectively retard the migration of one of the sequences compared to the other, effectively making it apparently larger and, thus, providing a means of differentiating between the two. The use of HA-red has been shown to resolve sequences with an AT variation of less than 1% (Wawer *et al.* 1995), whilst Hanse Analytik have reported that HA staining is so sensitive that in one case it was used to distinguish two 567bp sequences which differed by only a single point mutation (Hanse Analytik 1996, personal communication). Therefore, if one wished to check

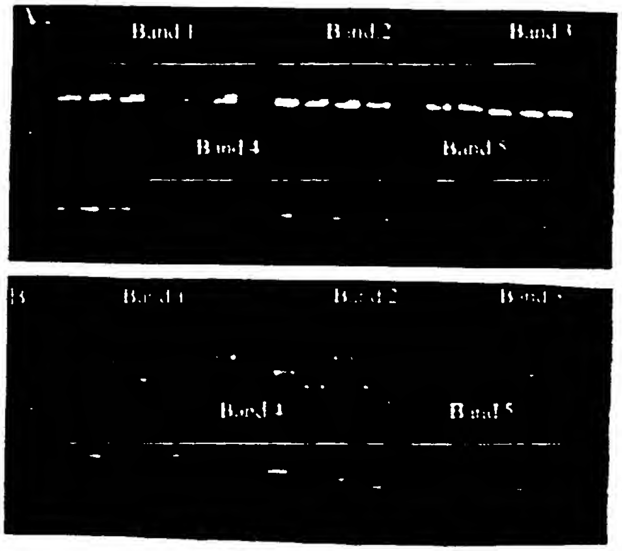


Figure 10 Discrimination of clones of identical/nearly identical size using HA-red. Bands of decreasing size (1–5) were extracted from the final display of a suppression subtractive hybridization experiment and cloned. Seven colonies were picked at random from each cloned band and their inserts amplified using PCR. The products were run on two gels, (A) a high resolution 2 % agarose gel, and (B) a high resolution 2 % agarose gel containing 1 U/ml HA-red. With few exceptions, all the clones from each band appear to be the same size (gel A). However, the presence of HA-red (gel B), which separates identically-sized DNA fragments based on the percentage of GC within the sequence, clearly indicates the presence of different gene species within each band. For example, even though all five re-amplified clones of band 1 appear to be the same size, at least four different gene species are represented.

in a similar gel containing one of the HA-stains. The standard gel should indicate any gross size differences, whilst the HA-stained gel should separate otherwise unresolvable species (on standard AGE) according to their base content. Geisinger *et al.* (1997) reported successful use of this approach for identifying DD-derived clones. Figure 10 shows such an experiment carried out in this laboratory on clones obtained from a band extracted from an SSH display.

An alternative approach is to carry out a 2-D analysis of the differential display products. In this approach, size-based separation is first carried out in a standard agarose gel. The gel slice containing the display is then extracted and incorporated in to a HA gel for resolution based on AT/GC content.

Of course, one should always consider the possibility of there being different gene species which are the same size and have the same GC/AT content. However, even these species are not unresolvable given some effort—again, one might use SSCP, or perhaps a denaturing gradient gel electrophoresis (DGGE) or temperature gradient field electrophoresis (TGGE) approach to resolve the contents of a band, either directly on the extracted band (Suzuki *et al.* 1991) or on the reamplified product.

The requirement of some differential display techniques to visualize large numbers of products (e.g. DD and GEF) can also present a problem in that, in terms of numbers, the resolution of PAGE rarely exceeds 300–400 bands. One approach to overcoming this might be to use 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991).

[A-red. Bands of decreasing
ı subtractive hybridization
each cloned band and their
high resolution 2 % agarose
-ed. With few exceptions, all
-er, the presence of HA-red
he percentage of GC within
-ies within each band. For
-e the same size, at least four

rd gel should indicate
ld separate otherwise
ase content. Geisinger
entifying DD-derived
is laboratory on clones

he differential display
ried out in a standard
cted and incorporated

there being different
AT content. However,
-again, one might use
)GGE) or temperature
he contents of a band,
or on the reamplified

ies to visualize large
oblem in that, in terms

Extraction of differentially expressed bands from a gel can be complex since, in some cases (e.g. DD, GEF), the results are visualized by autoradiographic means, such that precise overlay of the developed film on the gel must occur if the correct band is to be extracted for further analysis. Clearly, a misjudged extraction can account for many man-hours lost. This problem, and that of the use of radioisotopes, has been addressed by several groups. For example, Lohmann *et al.* (1995) demonstrated that silver staining can be used directly to visualize DD bands in horizontal PAGs. An *et al.* (1996) avoided the use of radioisotopes by transferring a small amount (20–30 %) of the DNA from their DD to a nylon membrane, and visualizing the bands using chemiluminescent staining before going back to extract the remaining DNA from the gel. Chen and Peck (1996) went one step further and transferred the entire DD to a nylon membrane. The DNA bands were then visualized using a digoxigenin (DIG) system (DIG was attached to the polydT primers used in the differential display procedure). Differentially expressed bands were cut from the membrane and the DNA eluted by washing with PCR buffer prior to reamplification.

One of the advantages of using techniques such as SSH and RDA is that the final display can be run on an agarose gel and the bands visualized with simple ethidium bromide staining. Whilst this approach can provide acceptable results, overstaining with SYBR Green I or SYBR Gold nucleic acid stains (FMC) effectively enhances the intensity and sharpness of the bands. This greatly aids in their precise extraction and often reveals some faint products that may otherwise be overlooked. Whilst differential displays stained with SYBR Green I are better visualized using short wavelength UV (254 nm) rather than medium wavelength (306 nm), the shorter wavelength is much more DNA damaging. In practice, it takes only a few seconds to damage DNA extracted under 254 nm irradiation, effectively preventing reamplification and cloning. The best approach is to overstain with SYBR Green I and extract bands under a medium wavelength UV transillumination.

The possible use of 'microfingerprinting' to reduce complexity

Given the sheer number of gene products and the possible complexity of each band, an alternative approach to rapid characterization may be to use an enhanced analysis of a small section of a differential display—a 'sub-fingerprint' or 'microfingerprint'. In this case, one could concentrate on those bands which only appear in a particular chosen size region. Reducing the fingerprint in this way has at least two advantages. One is that it should be possible to use different gel types, concentrations and run times tailored exactly to that region. Currently, one might run products from 100–3000 – bp on the same gel, which leads to compromize in the gel system being used and consequently to suboptimal resolution, both in terms of size and numbers, and can lead to problems in the accurate excision of individual bands. Secondly, it may be possible to enhance resolution by using a 2-D analysis using a HA-stain, as described earlier. In summary, if a range of gene product sizes is carefully chosen to included certain 'relevant' genes, the 2-D system standardized, and appropriate gene analysis used, it may be possible to develop a method for the early and rapid identification of compounds which have similar or widely different

An alternative approach to microfingerprinting is to examine altered expression in specific families of genes through careful selection of PCR primers and/or post-reaction analysis. Stress genes, growth factors and/or their receptors, cell cycling genes, cytochromes P450 and regulatory proteins might be considered as candidates for analysis in this way. Indeed, some off-the-shelf DNA arrays (e.g. Clontech's Atlas cDNA Expression Array series) already anticipated this to some degree by grouping together genes involved in different responses e.g. apoptosis, stress, DNA-damage response etc.

Screening

False positives

The generation of false positives has been discussed at length amongst the differential display community (Liang *et al.* 1993, 1995, Nishio *et al.* 1994, Sun *et al.* 1994, Sompayrac *et al.* 1995). The reason for false positives varies with the technique being used. For instance, in RDA, the use of adaptors which have not been HPLC purified can lead to the production of false positives through illegitimate ligation events (O'Neill and Sinclair 1997), whilst in DD they can arise through PCR artifacts and illegitemate transcription of rRNA. In SH, false positives appear to be derived largely from abundant gene species, although some may arise from cDNA/mRNA species which do not undergo hybridization for technical reasons.

A quick screening of putative differentially expressed clones can be carried out using a simple dot blot approach, in which labelled first strand probes synthesized from tester and driver mRNA are hybridized to an array of said clones (Hedrick *et al.* 1984, Sakaguchi *et al.* 1986). Differentially expressed clones will hybridize to tester probe, but not driver. The disadvantage of this approach is that rare species may not generate detectable hybridization signals. One option for those using SSH is to screen the clones using a labelled probe generated from the subtracted cDNA from which it was derived, and with a probe made from the reverse subtraction reaction (ClonTechniques 1997a). Since the SSH method enriches rare sequences, it should be possible to confirm the presence of clones representing low abundance genes. Despite this quick screening step, there is still the need to go back to the original mRNA and confirm the altered expression using a more quantitative approach. Although this may be achieved using Northern blots, the sensitivity is poor by today's high standards and one must rely on PCR methods for accurate and sensitive determinations (see below).

Sequence analysis

The majority of differential display procedures produce final products which are between 100 and 1000bp in size. However, this may considerably reduce the size of the sequence for analysis of the DNA databases. This in turn leads to a reduced confidence in the result—several families of genes have members whose DNA sequences are almost identical except in a few key stretches, e.g. the cytochrome P450 gene superfamily (Nelson *et al.* 1996). Thus, does the clone identified as being almost identical to gene X_0 really come from that gene, or its brother gene X_1 or its as yet undiscovered sister X_2? For example, using SSH, part of a gene was isolated,

mine altered expression
R primers and/or post-
: receptors, cell cycling
onsidered as candidates
arrays (e.g. Clontech's
this to some degree by
ipoptosis, stress, DNA-

at length amongst the
iio *et al.* 1994, Sun *et al.*
sitives varies with the
laptors which have not
ves through illegitimate
they can arise through
I, false positives appear
1 some may arise from
1 for technical reasons.
ones can be carried out
and probes synthesized
said clones (Hedrick *et*
lones will hybridize to
ach is that rare species
on for those using SSH
1 the subtracted cDNA
he reverse subtraction
nriches rare sequences,
senting low abundance
need to go back to the
: a more quantitative
blots, the sensitivity is
ethods for accurate and

nal products which are
rably reduce the size of
urn leads to a reduced
nembers whose DNA

which was up-regulated in the liver of rats exposed to Wy-14,643 and was identified by a FASTA search as being transferrin (data not shown). However, transferrin is known to be downregulated by hypolipidemic peroxisome proliferators such as Wy-14,643 (Hertz *et al.* 1996), and this was confirmed with subsequent RT-PCR analysis. This suggests that the gene sequence isolated may belong to a gene which is closely related to transferrin, but is regulated by a different mechanism.

A further problem associated with SH technology is redundancy. In most cases before SH is carried out, the cDNA population must first be simplified by restriction digestion. This is important for at least two reasons:

(1) To reduce complexity—long cDNA fragments may form complex networks which prevent the formation of appropriate hybrids, especially at the high concentrations required for efficient hybridization.
(2) Cutting the cDNAs into small fragments provides better representation of individual genes. This is because genes derived from related but distinct members of gene families often have similar coding sequences that may cross-hybridize and be eliminated during the subtraction procedure (Ko 1990). Furthermore, different fragments from the same cDNA may differ considerably in terms of hybridization and amplification and, thus, may not efficiently do one or the other (Wang and Brown 1991). Thus, some fragments from differentially expressed cDNAs may be eliminated during subtractive hybridization procedures. However, other fragments may be enriched and isolated. As a consequence of this, some genes will be cut one or more times, giving rise to two or more fragments of different sizes. If those same genes are differentially expressed, then two or more of the different size fragments may come through as separate bands on the final differential display, increasing the observed redundancy and increasing the number of redundant sequencing reactions.

Sequence comparisons also throw up another important point—at what degree of sequence similarity does one accept a result. Is 90% identity between a gene derived from your model species and another acceptably close? Is 95% between your sequence and one from the same species also acceptable? This problem is particularly relevant when the forward and reverse sequence comparisons give similar sequences with completely different gene species! An arbitrary decision seems to be to allocate genes that are definite (95% and above similarity) and then group those between 60 and 95% as being related or possible homologues.

Quantitative analysis

At some point, one must give consideration to the quantitative analysis of the candidate genes, either as a means of confirming that they are truly differentially expressed, or in order to establish just what the differences are. Northern blot analysis is a popular approach as it is relatively easy and quick to perform. However, the major drawback with Northern blots is that they are often not sensitive enough to detect rare sequences. Since the majority of messages expressed in a cell are of low abundance (see table 1), this is a major problem. Consequently, RT-PCR may be the method of choice for confirming differential expression. Alt appropriate thermal cycling technology. Whilst quantitative analysis is more desirable, being more accurate and without reliance on an internal standard, the money and time needed to develop a competitor molecule is often excessive, especially when one might be examining tens or even hundreds of gene species. The use of semi-quantitative analysis is simpler, although still relatively involved. One must first of all choose an internal standard that does not change in the test cells compared to the controls. Numerous reference genes have been tried in the past, for example interferon-gamma (IFN-γ, Frye *et al.* 1989), β-actin (Heuval *et al.* 1994), glyceraldehyde-3-phosphate dehydrogenase (GAPDH, Wong *et al.* 1994), dihydrofolate reductase (DHFR, Mohler and Butler 1991), β-2-microglobulin (β-2-m, Murphy *et al.* 1990), hypoxanthine phosphoribosyl transferase (HPRT, Foss *et al.* 1998) and a number of others (ClonTechniques 1997b). Ideally, an internal standard should not change its level of expression in the cell regardless of cell age, stage in the cell cycle or through the effects of external stimuli. However, it has been shown on numerous occasions that the levels of most housekeeping genes currently used by the research community do in fact change under certain conditions and in different tissues (ClonTechniques 1997b). It is imperative, therefore, that preliminary experiments be carried out on a panel of housekeeping genes to establish their suitability for use in the model system.

Interpretation of quantitative data must also be treated with caution. By comparing the lists of genes identified by differential expression one can perhaps gain insight into why two different species react in different ways to external stimuli. For example, rats and mice appear sensitive to the non-genotoxic effects of a wide range of peroxisome proliferators whilst Syrian hamsters and guinea pigs are largely resistant (Orton *et al.* 1984, Rodricks and Turnbull 1987, Lake *et al.* 1989, 1993, Makowska *et al.* 1992). A simplified approach to resolving the reason(s) why is to compare lists of up- and down-regulated genes in order to identify those which are expressed in only one species and, through background knowledge of the effects of the said gene, might suggest a mechanism of facilitated non-genotoxic carcinogenesis or protection. Of course, the situation is likely to be far more complex. Perhaps if there were one key gene protecting guinea pig from non-genotoxic effects and it was upregulated 50 times by PPs, the same gene might only be up-regulated five times in the rat. However, since both were noted to be upregulated, the importance of the gene may be overlooked. Just to complicate matters, a large change in expression does not necessarily mean a biologically important change. For example, what is the true relevance of gene Y which shows a 50-fold increase after a particular treatment, and gene Z which shows only a 5-fold increase? If one examines the literature one may find that historically, gene Y has often been shown to be up-regulated 40–60-fold by a number of unrelated stimuli—in light of this the 50-fold increase would appear less significant. However, the literature may show that gene Z has never been recorded as having more than doubled in expression—which makes your 5-fold increase all the more exciting. Perhaps even more interesting is if that same 5-fold increase has only been seen in related neoplasms or following treatment with related chemicals.

Problems in using the differential display approach

Differential display technology originally held promise of an easily obtainable 'fingerprint' of those genes which are up- or down-regulated in test animals/cells in a developmental process or following exposure to given stimuli. However, it has

ative analysis is more
internal standard, the
cule is often excessive,
eds of gene species. The
elatively involved. One
change in the test cells
een tried in the past, for
in (Heuval *et al.* 1994),
Vong *et al.* 1994), di-
β-2-microglobulin (β-2-
sferase (HPRT, Foss *et*
b). Ideally, an internal
ll regardless of cell age,
li. However, it has been
keeping genes currently
ertain conditions and in
e, therefore, that pre-
eping genes to establish

ated with caution. By
ession one can perhaps
ways to external stimuli.
otoxic effects of a wide
d guinea pigs are largely
Lake *et al.* 1989, 1993,
the reason(s) why is to
dentify those which are
wledge of the effects of
enotoxic carcinogenesis
re complex. Perhaps if
otoxic effects and it was
up-regulated five times
d, the importance of the
e change in expression
or example, what is the
a particular treatment,
mines the literature one
be up-regulated 40–60-
50-fold increase would
at gene Z has never been
hich makes your 5-fold
ng is if that same 5-fold
g treatment with related

However, it has

become clear that the fingerprinting process, whilst still valid, is much too complex to be represented by a single technique profile. This is because all differential display techniques have common and/or unique technical problems which preclude the isolation and identification of all those genes which show changes in expression. Furthermore, there are important genetic changes related to disease development which differential expression analysis is simply not designed to address. An example of this is the presence of small deletions, insertions, or point mutations such as those seen in activated oncogenes, tumour suppressor genes and individual polymorphisms. Polymorphic variations, small though they usually are, are often regarded as being of paramount importance in explaining why some patients respond better than others to certain drug treatments (and, in logical extension, why some people are less affected by potentially dangerous xenobiotics/carcinogens than others). The identification of such point mutations and naturally occurring polymorphisms requires the subsequent application of sequencing, SSCP, DGGE or TGGE to the gene of interest. Furthermore, differential display is not designed to address issues such as alternatively spliced gene species or whether an increased abundance of mRNA is a result of increased transcription or increased mRNA stability.

Conclusions

Perhaps the main advantage of open system differential display techniques is that they are not limited by extant theories or researcher bias in revealing genes which are differentially expressed, since they are designed to amplify all genes which demonstrate altered expression. This means that they are useful for the isolation of previously unknown genes which may turn out be useful biomarkers of a particular state or condition. At least one open system (SAGE) is also quantitative, thus eliminating the need to return to the original mRNA and carry out Northern/PCR analysis to confirm the result. However, the rapid progress of genome mapping projects means that over the next 5–10 years or so, the balance of experimental use will switch from open to closed differential display systems, particularly DNA arrays. Arrays are easier and faster to prepare and use, provide quantitative data, are suitable for high throughput analysis and can be tailored to look at specific signalling pathways or families of genes. Identification of all the gene sequences in human and common laboratory animals combined with improved DNA array technology, means that it will soon no longer be necessary to try to isolate differentially expressed genes using the technically more demanding open system approach. Thus, their main advantage (that of identifying unknown genes) will be largely eradicated. It is likely, therefore, that their sphere of application will be reduced to analysis of the less common laboratory species, since it will be some time yet before the genomes of such animals as zebrafish, electric eels, gerbils, crayfish and squid, for example, will be sequenced.

Of course, in the end the question will always remain: What is the functional/biological significance of the identified, differentially expressed genes? One

replication will also be upregulated but may have little or nothing to do with the carcinogenic effect. Whilst differential display technology cannot hope to answer these questions, it does provide a springboard from which identification, regulatory and functional studies can be launched. Understanding the molecular mechanism of cellular responses is almost impossible without knowing the regulation and function of those genes and their condition (e.g. mutated). In an abstract sense, differential display can be likened to a still photograph, showing details of a fixed moment in time. Consider the Historian who knows the outcome of a battle and the placement and condition of the troops before the battle commenced, but is asked to try and deduce how the battle progressed and why it ended as it did from a few still photographs—an impossible task. In order to understand the battle, the Historian must find out the capabilities and motivation of the soldiers and their commanding officers, what the orders were and whether they were obeyed. He must examine the terrain, the remains of the battle and consider the effects the prevailing weather conditions exerted. Likewise, if mechanistic answers are to be forthcoming, the scientist must use differential display in combination with other techniques, such as knockout technology, the analysis of cell signalling pathways, mutation analysis and time and dose response analyses. Although this review has emphasized the importance of differential gene profiling, it should not be considered in isolation and the full impact of this approach will be strengthened if used in combination with functional genomics and proteomics (2-dimensional protein gels from isoelectric focusing and subsequent SDS electrophoresis and virtual 2D-maps using capillary electrophoresis). Proteomics is attracting much recent attention as many of the changes resulting in differential gene expression do not involve changes in mRNA levels, as decribed extensively herein, but rather protein–protein, protein–DNA and protein phosphorylation events which would require functional genomics or proteomic technologies for investigation.

Despite the limitations of differential display technology, it is clear that many potential applications and benefits can be obtained from characterizing the genetic changes that occur in a cell during normal and disease development and in response to chemical or biological insult. In light of functional data, such profiling will provide a 'fingerprint' of each stage of development or response, and in the long term should help in the elucidation of specific and sensitive biomarkers for different types of chemical/biological exposure and disease states. The potential medical and therapeutic benefits of understanding such molecular changes are almost immeasurable. Amongst other things, such fingerprints could indicate the family or even specific type of chemical an individual has been exposed to plus the length and/or acuteness of that exposure, thus indicating the most prudent treatment. They may also help uncover differences in histologically identical cancers, provide diagnostic tests for the earliest stages of neoplasia and, again, perhaps indicate the most efficacious treatment.

The Human Genome Project will be completed early in the next century and the DNA sequence of all the human genes will be known. The continuing development and evolution of differential gene expression technology will ensure that this knowledge contributes fully to the understanding of human disease processes.

Acknowledgements

We acknowledge Drs Nick Plant (University of Surrey), Sally Darney and Chris Luft (US EPA at RTP) for their critical analysis of the manuscript prior to submission. This manuscript has been reviewed in accordance with the policy of the

cannot hope to answer
.entification, regulatory
nolecular mechanism of
regulation and function
tract sense, differential
s of a fixed moment in
attle and the placement
but is asked to try and
it did from a few still
ne battle, the Historian
and their commanding
l. He must examine the
the prevailing weather
o be forthcoming, the
her techniques, such as
, mutation analysis and
has emphasized the
sidered in isolation and
d in combination with
n gels from isoelectric
D-maps using capillary
ention as many of the
olve changes in mRNA
tein, protein–DNA and
nctional genomics or

y, it is clear that many
racterizing the genetic
pment and in response
ta, such profiling will
ponse, and in the long
iomarkers for different
e potential medical and
anges are almost im-
indicate the family or
sed to plus the length
ost prudent treatment.
ntical cancers, provide
n, perhaps indicate the

he next century and the
ontinuing development
will ensure that this
disease processes.

US Environmental Protection Agency and approved for publication. Approval does not signify that the contents reflect the views and policies of the Agency, nor does mention of trade names constitute endorsement or recommendation for use.

References

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R. and Ventor, J. C., 1991, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

An, G., Luo, G., Veltri, R. W. and O'Hara, S. M., 1996, Sensitive non-radioactive differential display method using chemiluminescent detection. *Biotechniques*, **20**, 342–346.

Axel, R., Feigelson, P. and Schultz, G., 1976, Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell*, **7**, 247–254.

Band, V. and Sager, R., 1989, Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a medium that supports long-term growth of both cell types. *Proceedings of the Naional Academy of Sciences, USA*, **86**, 1249–1253.

Bauer, D., Muller, H., Reich, J., Riedel, H., Ahrenkiel, V., Warthoe, P. and Strauss, M., 1993, Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, **21**, 4272–4280.

Bertioli, D. J., Schlichter, U. H. A., Adams, M. J., Burrows, P. R., Steinbiss, H.-H. and Antoniw, J. F., 1995, An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Research*, **23**, 4520–4523.

Bravo, R., 1990, Genes induced during the G0/G1 transition in mouse fibroblasts. *Seminars in Cancer Biology*, **1**, 37–46.

Burn, T. C., Petrovick, M. S., Hohaus, S., Rollins, B. J. and Tenen, D. G., 1994, Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. *Blood*, **84**, 2776–2783.

Cao, J., Cai, X., Zheng, L., Geng, L., Shi, Z., Pao, C. C. and Zheng, S., 1997, Characterisation of colorectal cancer-related cDNA clones obtained by subtractive hybridisation screening. *Journal of Cancer Research and Clinical Oncology*, **123**, 447–451.

Cassidy, S. B., 1995, Uniparental disomy and genomic imprinting as causes of human genetic disease. *Environmental and Molecular Mutagenesis*, **25** (Suppl 26), 13–20.

Chang, G. W. and Terzaghi-Howe, M., 1998, Multiple changes in gene expression are associated with normal cell-induced modulation of the neoplastic phenotype. *Cancer Research*, **58**, 4445–4452.

Chen, J., Schwartz, D. A., Young, T. A., Norris, J. S. and Yager, J. D., 1996, Identification of genes whose expression is altered during mitosuppression in livers of ethinyl estradiol-treated female rats. *Carcinogenesis*, **17**, 2783–2786.

Chen, J. J. W. and Peck, K., 1996, Non-radioactive differential display method to directly visualise and amplify differential bands on nylon membrane. *Nucleic Acid Research*, **24**, 793–794

ClonTechniques, 1997a, PCR-Select Differential Screening Kit—the nextstep after Clontech PCR-Select cDNA subtraction. *ClonTechniques*, **XII**, 18–19

ClonTechniques, 1997b, Housekeeping RT-PCR amplimers and cDNA probes. *ClonTecnniques*, **XII**, 15–16

Davis, M. M., Cohen, D. I., Nielsen, E. A., Steinmetz, M., Paul, W. E. and Hood, L., 1984, Cell-type-specific cDNA probes and the murine I region: the localization and orientation of Ad alpha *Proceedings of the National Academy of Sciences (USA)*, **81**, 2194–2198

Dellavalle, R. P., Peterson, R. and Lindquist, S., 1994, Preferential deadenylation of HSP70 mRNA plays a key role in regulating Hsp70 expression in Drosophila melanogaster. *Molecular and Cell Biology*, **14**, 3646–3659.

DeRisi, J. L., Vashwanath, R. L. and Brown, P., 1997, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Diatchenko, L., Lau, Y.-F. C., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. and Siebert, P. D., 1996, Suppression subtractive hybridisation: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences (USA)*, **93**, 6025–6030.

Dogra, S. C., Whitelaw, M. L. and May, B. K., 1998, Transcriptional activation of cytochrome P450 genes by different classes of chemical inducers. *Clinical and Experimental Pharmacology and Physiology*, **25**, 1–9

FITZPATRICK, D. R., GERMAIN-LEE, E. and VALLE, D., 1995, Isolation and characterisation of rat and human cDNAs encoding a novel putative peroxisomal enoyl-CoA hydratase. *Genomics*, **27**, 457–466.

FOSS, D. L., BAARSCH, M. J. and MURTAUGH, M. P., 1998, Regulation of hypoxanthine phosphoribosyltransferase, glyceraldehyde-3-phosphate dehydrogenase and beta-actin mRNA expression in porcine immune cells and tissues. *Animal Biotechnology*, 9, 67–78

FRYE, R. A., BENZ, C. C. and LIU, E., 1989, Detection of amplified oncogenes by differential polymerase chain reaction. *Oncogene*, **4**, 1153–1157.

GEISINGER, A., RODRIGUEZ, R., ROMERO, V and WETTSTEIN R., 1997, A simple method for screening cDNAs arising from the cloning of RNA differential display bands. *Elsevier Trends Journals Technical Tips Online*, http://tto.trends.com, document T01110.

GRESS, T. M., HOHEISEL, J. D., LENNON, G. G., ZEHETNER, G. and LEHRACH, H., 1992, Hybridisation fingerprinting of high density cDNA filter arrays with cDNA pools derived from whole tissues. *Mammalian Genome*, **3**, 609–619.

GRIFFIN, G. and KRISHNA, S., 1998, Cytokines in infectious diseases. *Journal of the Royal College of Physicians, London*, **32**, 195–198.

GROENINK, M. and LEEGWATER, A. C. J., 1996, Isolation of delayed early genes associated with liver regeneration using Clontech PCR-select subtraction technique. *Clontechniques*, **XI**, 23–24

GUIMARAES, M. J., BAZAN, J. F., ZLOTNIK, A., WILES, M. V., GRIMALDI, J. C., LEE, F. and MCCLANAHAN, T., 1995b, A new approach to the study of haematopoietic development in the yolk sac and embryoid bodies. *Development*, 121, 3335–3346.

GUIMERAES, M. J., LEE, F., ZLOTNIK, A. and MCCLANAHAN, T., 1995a, Differential display by PCR: novel findings and applications. *Nucleic Acids Research*, **23**, 1832–1833.

GURSKAYA, N. G., DIATCHENKO, L., CHENCHIK, P. D., SIEBERT, P. D., KHASPEKOV, G. L., LUKYANOV, K. A., VAGNER, L. L., ERMOLAEVA, O. D., LUKYANOV, S. A. and SVERDLOV, E. D., 1996, Equalising cDNA subtraction based on selective suppression of polymerase chain reaction: Cloning of Jurkat cell transcripts induced by phytohemaglutinin and phorbol 12-Myrystate 13-Acetate. *Analytical Biochemistry*, **240**, 90–97.

HAMPSON, I. N. and HAMPSON, L., 1997, CCLS and DROP—subtractive cloning made easy. *Life Science News* (A publication of Amersham Life Science), **23**, 22–24.

HAMPSON, I. N., HAMPSON, L. and DEXTER, T. M., 1996, Directional random oligonucleotide primed (DROP) global amplification of cDNA: its application to subtractive cDNA cloning. *Nucleic Acids Research*, **24**, 4832–4835.

HAMPSON, I. N., POPE, L., COWLING, G. J. and DEXTER, T. M., 1992, Chemical cross linking subtraction (CCLS): a new method for the generation of subtractive hybridisation probes. *Nucleic Acids Research*, **20**, 2899.

HARA, E., KATO, T., NAKADA, S., SEKIYA, S. and ODA, K., 1991, Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. *Nucleic Acids Research*, **19**, 7097–7104.

HATADA, I., HAYASHIZAKE, Y., HIROTSUNE, S., KOMATSUBARA, H. and MUKAI, T., 1991, A genomic scanning method for higher organisms using restriction sites as landmarks. *Proceedings of the National Academy of Sciences (USA)*, **88**, 9523–9527.

HECHT, N., 1998, Molecular mechanisms of male sperm cell differentiation. *Bioessays*, **20**, 555–561.

HEDRICK, S., COHEN, D. I., NIELSEN, E. A. and DAVIS, M. E., 1984, Isolation of T cell-specific membrane-associated proteins. *Nature*, **308**, 149–153.

HERTZ, R., SECKBACH, M., ZAKIN, M. M. and BAR-TANA, J., 1996, Transcriptional suppression of the transferrin gene by hypolipidemic peroxisome proliferators. *Journal of Biological Chemistry*, **271**, 218–224.

HEUVAL, J. P. V., CLARK, G. C., KOHN, M. C., TRITSCHER, A. M., GREENLEE, W. F., LUCIER, G. W. and BELL, D. A., 1994, Dioxin-responsive genes: Examination of dose-response relationships using quantitative reverse transciptase-polymerase chain reaction. *Cancer Research*, **54**, 62–68.

HILLIER, L. D., LENNON, G., BECKER, M., BONALDO, M. F., CHIAPELLI, B., CHISSOE, S., DIETRICH, N., DUBUQUE, T., FAVELLO, A., GISH, W., HAWKINS, M., HULTMAN, M., KUCABA, T., LACY, M., LE, M., LE, N., MARDIS, E., MOORE, B., MORRIS, M., PARSONS, J., PRANGE, C., RIFKIN, L., ROHLFING, T., SCHELLENBERG, K., SOARES, M. B., TAN, F., THIERRY-MEG, J., TREVASKIS, E., UNDERWOOD, K., WOHLDMAN, P., WATERSTON, R., WILSON, R and MARRA, M., 1996, Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, **6**, 807–828.

HUBANK, M. and SCHATZ, D. G., 1994, Identifying differences in mRNA expression by representational difference analysis. *Nucleic Acids Research*, **22**, 5640–5648.

HUNTER, T., 1991, Cooperation between oncogenes. *Cell*, **64**, 249–270.

IVANOVA, N. B. and BELYAVSKY, A. V., 1995, Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Research*, **23**, 2954–2958.

JAMES, B. D. and HIGGINS, S. J, 1985, *Nucleic Acid Hybridisation* (Oxford: IRL Press Ltd).

KAS-DEELEN, A. M., HARMSEN, M. C., DE MAAR, E. F. and VAN SON, W. J, 1998, A sensitive method for

ınd characterisation of rat and
CoA hydratase. *Genomics*, 27,

ın of hypoxanthine phospho-
d beta-actin mRNA expression
78

enes by differential polymerase

A simple method for screening
ınds. *Elsevier Trends Journals*

RACH, H., 1992, Hybridisation
ols derived from whole tissues.

Journal of the Royal College of

rly genes associated with liver
Clontechniques, **XI**, 23–24.
RIMALDI, J. C., LEE, F and
poietic development in the yolk

995a, Differential display by
1832–1833.
KHASPEKOV, G. L., LUKYANOV,
and SVERDLOV, E. D., 1996,
of polymerase chain reaction:
and phorbol 12-Myrystate 13-

cloning made easy. *Life Science*

andom oligonucleotide primed
active cDNA cloning. *Nucleic*

mical cross linking subtraction
disation probes. *Nucleic Acids*

tractive cDNA cloning using
c to undifferentiated human
4
MUKAI, T., 1991, A genomic
landmarks. *Proceedings of the*

on. *Bioessays*, **20**, 555–561.
. Isolation of T cell-specific

scriptional suppression of the
l of *Biological Chemistry*, **271**.

LEE, W. F., LUCIER, G. W. and
e-response relationships using
er *Research*, **54**, 62–68.
B., CHISSOE, S., DIETRICH, N.,
I., KUCABA, T., LACY, M., LE,
GE, C., RIFKIN, L., ROHLFING,
TREVASKIS, E., UNDERWOOD,
1996, Generation and analysis
07–828.
expression by representational

expressed genes by restricti

quantifying cytomegalic endothelial cells in peripheral blood from cytomegalovirus-infected patients. *Clinical Diagnostic and Laboratory Immunology*, **5**, 622–626.

KILTY, I. and VICKERS, P., 1997, Fractionating DNA fragments generated by differential display PCR. *Strategies Newsletter* (Stratagene), **10**, 50–51.

KLEINJAN, D.-J. and VAN HEYNINGEN, V., 1998, Position effect in human genetic disease. *Human and Molecular Genetics*, **7**, 1611–1618.

KO, M. S., 1990, An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research*, **18**, 5705–5711.

LAKE, B. G., EVANS, J. G., CUNNINGHAME, M. E. and PRICE, R. J., 1993, Comparison of the hepatic effects of Wy-14,643 on peroxisome proliferation and cell replication in the rat and Syrian hamster. *Environmental Health Perspectives*, **101**, 241–248.

LAKE, B. G., EVANS, J. G., GRAY, T. J. B., KOROSI, S. A. and NORTH, C. J., 1989, Comparative studies of nafenopin-induced hepatic peroxisome proliferation in the rat, Syrian hamster, guiea pig and marmoset. *Toxicology and Applied Pharmacology*, **99**, 148–160.

LENNARD, M. S., 1993, Genetically determined adverse drug reactions involving metabolism. *Drug Safety*, **9**, 60–77.

LEVY, S., TODD, S. C. and MAECKER, H. T., 1998, CD81(TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system. *Annual Review of Immunology*, **16**, 89–109.

LIANG, P. and PARDEE, A. B., 1992, Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967–971.

LIANG, P., AVERBOUKH, L., KEYOMARSI, K., SAGER, R. and PARDEE, A., 1992, Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Research*, **52**, 6966–6968.

LIANG, P., AVERBOUKH, L. and PARDEE, A. B., 1993, Distribution & cloning of eukaryotic mRNAs by means of differential display refinements and optimisation. *Nucleic Acids Research*, **21**, 3269–3275.

LIANG, P., BAUER, D., AVERBOUKH, L., WARTHOE, P., ROHRWILD, M., MULLER, H., STRAUSS, M. and PARDEE, A. B., 1995, Analysis of altered gene expression by differential display. *Methods in Enzymology*, **254**, 304–321.

LINSKENS, M. H., FENG, J., ANDREWS, W. H., ENLOW, B. E., SAATI, S. M., TONKIN, L. A., FUNK, W. D. and VILLEPONTEAU, B., 1995, Cataloging altered gene expression in young and senescent cells using enhanced differential display. *Nucleic Acids Research*, **23**, 3244–3251.

LISITSYN, N., LISITSYN, N. and WIGLER, M., 1993, Cloning the differences between two complex genomes. *Science*, **259**, 946–951.

LOHMANN, J., SCHICKLE, H. and BOSCH, T. C. G., 1995, REN Display, a rapid and efficient method for non-radioactive differential display and mRNA isolation. *Biotechniques*, **18**, 200–202.

LUNNEY, J. K., 1998, Cytokines orchestrating the immune response. *Reviews in Science and Techology*, **17**, 84–94.

MAKOWSKA, J. M., GIBSON, G. G. and BONNER, F. W., 1992, Species differences in ciprofibrate-induction of hepaic cytochrome P4504A1 and peroxisome proliferation. *Journal of Biochemical Toxicology*, **7**, 183–191.

MALDARELLI, F., XIANG, C., CHAMOUN, G. and ZEICHNER, S. L., 1998, The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. *Virus Research*, **53**, 39–51.

MATHIEU-DAUDE, F., CHENG, R., WELSH, J. and MCCLELLAND, M., 1996, Screening of differentially amplified cDNA products from RNA arbitrarily primed PCR fingerprints using single strand conformation polymorphism (SSCP) gels. *Nucleic Acids Research*, **24**, 1504–1507.

MCKENZIE, D. and DRAKE, D., 1997, Identification of differentially expressed gene products with the castaway system. *Strategies Newsletter* (Stratagene), **10**, 19–20.

MCCLELLAND, M., MATHIEU-DAUDE, F. and WELSH, J., 1996, RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends in Genetics*, **11**, 242–246.

MECHLER, B. and RABBITTS, T. H., 1981, Membrane-bound ribosomes of myeloma cells. IV. mRNA complexity of free and membrane-bound polysomes. *Journal of Cell Biology*, **88**, 29–36.

MEYER, U. A. and ZANGER, U. M., 1997, Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annual Review of Pharmacology and Toxicology*, **37**, 269–296.

MOHLER, K. M. and BUTLER, L. D., 1991, Quantitation of cytokine mRNA levels utilizing the reverse transcriptase-polymerase chain reaction following primary antigen-specific sensitization in vivo—I. Verification of linearity, reproducibility and specificity. *Molecular Immunology*, **28**, 437–447.

MURPHY, L. D., HERZOG, C. E., RUDICK, J. B., TITO FOJO, A. and BATES, S. E., 1990, Use of the polymerase chain reaction in the quantitation of the mdr-1 gene expression. *Biochemistry*, **29**, 10351–10356.

Nishio, Y., Aiello, L. P. and King, G. L., 1994, Glucose induced genes in bovine aortic smooth muscle cells identified by mRNA differential display. *FASEB Journal*, **8**, 103–106.

O'Neill, M. J. and Sinclair, A. H., 1997, Isolation of rare transcripts by representational difference analysis. *Nucleic Acids Research*, **25**, 2681–2682.

Orton, T. C., Adam, H. K., Bentley, M., Holloway, B. and Tucker, M. J., 1984, Clobuzarit: species differences in the morphological and biochemical response of the liver following chronic administration. *Toxicology and Applied Pharmacology*, **73**, 138–151.

Pelkonen, O., Maenpaa, J., Taavitsainen, P., Rautio, A. and Raunio, H., 1998, Inhibition and Induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica*, **28**, 1203–1253.

Philips, S. M., Bendall, A. J. and Ramshaw, I. A., 1990, Isolation of genes associated with high metastatic potential in rat mammary adenocarcinomas. *Journal of the National Cancer Institute*, **82**, 199–203.

Prashar, Y. and Weissman, S. M., 1996, Analysis of differential gene expression by display of 3'end restriction fragments of cDNAs. *Proceedings of the National Academy of Sciences (USA)*, **93**, 659–663.

Ragno, S., Estrada, I., Butler, R. and Colston, M. J., 1997, Regulation of macrophage gene expression following invasion by *Mycobacterium tuberculosis*. *Immunology Letters*, **57**, 143–146.

Ramana, K. V. and Kohli, K. K., 1998, Gene regulation of cytochrome P450—an overview. *Indian Journal of Experimental Biology*, **36**, 437–446.

Richard, L., Velasco, P. and Detmar, M., 1998, A simple immunomagnetic protocol for the selective isolation and long-term culture of human dermal microvascular endothelial cells. *Experimental Cell Research*, **240**, 1–6.

Rockett, J. C., Esdaile, D. J. and Gibson, G. G., 1997, Molecular profiling of non-genotoxic hepatocarcinogenesis using differential display reverse transcription-polymerase chain reaction (ddRT-PCR). *European Journal of Drug Metabolism and Pharmacokinetics*, **22**, 329–333.

Rodricks, J. V. and Turnbull, D., 1987, Inter-species differences in peroxisomes and peroxisome proliferation. *Toxicology and Industrial Health*, **3**, 197–212.

Rogler, G., Hausmann, M., Vogl, D., Aschenbrenner, E., Andus, T., Falk, W., Andreesen, R., Scholmerich, J. and Gross, V., 1998, Isolation and phenotypic characterization of colonic macrophages. *Clinical and Experimental Immunology*, **112**, 205–215.

Rohn, W. M., Lee, Y. J. and Benveniste, E. N., 1996, Regulation of class II MHC expression. *Critical Reviews in Immunology*, **16**, 311–330.

Rudin, C. M. and Thompson, C. B., 1998, B-cell development and maturation. *Seminars in Oncology*, **25**, 435–446.

Sakaguchi, N., Berger, C. N. and Melchers, F., 1986, Isolation of a cDNA copy of an RNA species expressed in murine pre-B cells. *EMBO Journal*, **5**, 2139–2147.

Sambrook, J., Fritsch, E. F. and Maniatis, T., 1989, Gel electrophoresis of DNA. In N. Ford, M. Nolan and M. Fergusen (eds), *Molecular Cloning—A laboratory manual*, 2nd edition (New York: Cold Spring Harbour Laboratory Press), Volume 1, pp. 6–37.

Sargent, T. D. and Dawid, I. B., 1983, Differential gene expression in the gastrula of Xenopus laevis. *Science*, **222**, 135–139.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown., P. O. and Davis, R. W., 1996, Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences (USA)*, **93**, 10614–10619.

Schneider, C., King, R. M. and Philipson, L., 1988, Genes specifically expressed at growth arrest of mammalian cells. *Cell*, **54**, 787–793.

Schneider-Maunoury, S., Gilardi-Hebenstreit, P. and Charnay, P., 1998, How to build a vertebrate hindbrain. Lessons from genetics. *C R Academy of Science III*, **321**, 819–834.

Semenza, G. L., 1994, Transcriptional regulation of gene expression: mechanisms and pathophysiology. *Human Mutations*, **3**, 180–199.

Sewall, C. H., Bell, D. A., Clark, G. C., Tritscher, A. M., Tully, D. B., Vanden Heuvel, J. and Lucier, G. W., 1995, Induced gene transcription: implications for biomarkers. *Clinical Chemistry*, **41**, 1829–1834.

Singh, N., Agrawal, S. and Rastogi, A. K., 1997, Infectious diseases and immunity: special reference to major histocompatibility complex. *Emerging Infectious Diseases*, **3**, 41–49.

Smith, N. R., Li, A., Aldersley, M., High, A. S., Markham, A. F. and Robinson, P. A., 1997, Rapid determination of the complexity of cDNA bands extracted from DDRT-PCR polyacrylamide gels. *Nucleic Acids* Research, **25** , 3552–3554.

Sompayrac, L., Jane, S., Burn., T. C., Tenen, D. G. and Danna, K. J., 1995, Overcoming limitations of the mRNA differential display technique. *Nucleic Acids Research*, **23**, 4738–4739.

St John, T. P. and Davis, R. W., 1979, Isolation of galactose-inducible DNA sequences from Saccharomyces cerevisiae by differential plaque filter hybridisation. *Cell*, **16**, 443–452.

Sun, Y., Hegamyer, G. and Colburn, N. H., 1994, Molecular cloning of five messenger RNAs differentially expressed in preneoplastic or neoplastic JB6 mouse epidermal cells: one is homologous to human tissue inhibitor of metalloproteinases-3. *Cancer Research*, **54**, 1139–1144.

s in bovine aortic smooth muscle
8, 103–106.

:s by representational difference

M. J., 1984, Clobuzarit: species
of the liver following chronic
51.

NIO, H., 1998, Inhibition and
tica, 28, 1203–1253.

of genes associated with high
of the National Cancer Institute,

expression by display of 3'end
lcademy of Sciences (USA), 93,

egulation of macrophage gene
munology Letters, 57, 143–146.

me P450—an overview. Indian

gnetic protocol for the selective
endothelial cells. Experimental

ar profiling of non-genotoxic
tion-polymerase chain reaction
acokinetics, 22, 329–333.

peroxisomes and peroxisome

T., Falk, W., Andreesen, R.,
pic characterization of colonic
15.

ss II MHC expression. Critical

turation. Seminars in Oncology,

DNA copy of an RNA species

resis of DNA. In N. Ford, M.
anual, 2nd edition (New York:

the gastrula of Xenopus laevis.

s, R. W., 1996, Parallel human
000 genes. Proceedings of the

expressed at growth arrest of

998, How to build a vertebrate
21, 819–834.

hanisms and pathophysiology

D. B., Vanden Heuvel, J. and
ons for biomarkers. Clinical

d immunity: special reference
3, 41–49.

Robinson, P. A., 1997, Rapid
DDRT-PCR polyacrylamide

1995, Overcoming limitations
ch, 23, 4738–4739

ucible DNA sequences from
n Cell 16 443–452

Sung, Y. J. and Denman, R. B., 1997, Use of two reverse transcriptases eliminates false-positive results in differential display. *Biotechniques*, **23**, 462–464

Sutton, G., White, O., Adams, M. and Kerlavage, A., 1995, TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1, 9–19

Suzuki, Y., Sekiya, T. and Hayashi, K., 1991, Allele-specific polymerase chain reaction: a method for amplification and sequence determination of a single component among a mixture of sequence variants. *Analytical Biochemistry*, **192**, 82–84.

Syed, V., Gu, W. and Hecht, N. B., 1997, Sertoli cells in culture and mRNA differential display provide a sensitive early warning assay system to detect changes induced by xenobiotics. *Journal of Andrology*, **18**, 264–273.

Uitterlinden, A. G., Slagboom, P., Knook, D. L. and Vijgl, J., 1989, Two-dimensional DNA fingerprinting of human individuals. *Proceedings of the National Academy of Sciences (USA)*, **86**, 2742–2746.

Ullman, K. S., Northrop, J. P., Verweij, C. L. and Crabtree, G. R., 1990, Transmission of signals from the T lymphocyte antigen receptor to the genes responsible for cell proliferation and immune function: the missing link. *Annual Review of Immunology*, **8**, 421–452.

Vasmatzis, G., Essand, M., Brinkmann, U., Lee, B. and Paston, I., 1998, Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proceedings of the National Academy of Sciences (USA)*, **95**, 300–304.

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W., 1995, Serial analysis of gene expression. *Science*, **270**, 484–487.

Voeltz, G. K. and Steitz, J. A., 1998, AuuuA sequences direct mRNA deadenylation uncoupled from decay during Xenopus early development. *Molecular and Cell Biology*, **18**, 7537–7545.

Vogelstein, B. and Kinzler, K. W., 1993, The multistep nature of cancer. *Trends in Genetics*, **9**, 138–141.

Walter, J., Belfield, M., Hampson, I. and Read, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, **21**, 13–14.

Wan, J. S., Sharp, S. J., Poirier, G. M.-C., Wagaman, P. C., Chambers, J., Pyati, J., Hom, Y.-L., Galindo, J. E., Huvar, A., Peterson, P. A., Jackson, M. R. and Erlander, M. G., 1996, Cloning differentially expressed mRNAs. *Nature Biotechnology*, **14**, 1685–1691.

Walter, J., Belfield, M., Hampson, I. and Read, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS, *Life Science News*, **21**, 13–14.

Wang, Z. and Brown, D. D. 1991, A gene expression screen. *Proceedings of the National Academy of Sciences (USA)*, **88**, 11505–11509.

Wawer, C., Ruggeberg, H., Meyer, G. and Muyzer, G., 1995, A simple and rapid electrophoresis method to detect sequence variation in PCR-amplified DNA fragments. *Nucleic Acids Research*, **23**, 4928–4929.

Welsh, J., Chada, K., Dalal, S. S., Cheng, R., Ralph, D. and McClelland, M., 1992, Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research*, **20**, 4965–4970

Wong, H., Anderson, W. D., Cheng, T. and Riabowol, K. T., 1994, Monitoring mRNA expression by polymerase chain reaction: the 'primer-dropping' method. *Analytical Biochemistry*, **223**, 251–258.

Wong, K. K. and McClelland, M., 1994, Stress-inducible gene of Salmonella typhimurium identified by arbitrarily primed PCR of RNA. *Proceedings of the National Academy of Sciences (USA)*, **91**, 639–643.

Wynford-Thomas, D., 1991, Oncogenes and anti-oncogenes; the molecular basis of tumour behaviour. *Journal of Pathology*, **165**, 187–201.

Xhu, D., Chan, W. L., Leung, B. P., Huang, F. P., Wheeler, R., Piedrafita, D., Robinson, J. H. and Liew, F. Y., 1998, Selective expression of a stable cell surface molecule on type 2 but not type 1 helper T cells. *Journal of Experimental Medicine*, **187**, 787–794

Yang, M. and Sytowski, A. J., 1996, Cloning differentially expressed genes by linker capture subtraction. *Analytical Biochemistry*, **237**, 109–114.

Zhao, N., Hashida, H., Takahashi, N., Misumi, Y. and Sakaki, Y., 1995, High-density cDNA filter analysis: a novel approach for large scale quantitative analysis of gene expression. *Gene*, **156**, 207–213.

Zhao, X. J., Newsome, J. T. and Cihlar, R. L., 1998, Up-regulation of two *candida albicans* genes in the rat model of oral candidiasis detected by differential display. *Microbial Pathogenesis*, **25**, 121–129.

Zimmermann, C. R., Orr, W. C., Leclerc, R. F., Barnard, C. and Timberlake, W. E., 1980, Molecular cloning and selection of genes regulated in *Aspergillus* development. *Cell*, **21**, 709–715.

Proc Natl Acad Sci USA
Vol 94, pp 8945–8947, August 1997
Applied Biological Sciences

Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR

DEVAL A. LASHKARI[*][†], JOHN H. MCCUSKER[‡], AND RONALD W. DAVIS[*][§]

[*]Departments of Genetics and Biochemistry, Beckman Center, Stanford University, Stanford, CA 94305, and [‡]Department of Microbiology, 3020 Duke University Medical Center, Durham, NC 27710

Contributed by Ronald W. Davis, May 20, 1997

ABSTRACT **The recent ability to sequence whole genomes allows ready access to all genetic material. The approaches outlined here allow automated analysis of sequence for the synthesis of optimal primers in an automated multiplex oligonucleotide synthesizer (AMOS). The efficiency is such that all ORFs for an organism can be amplified by PCR. The resulting amplicons can be used directly in the construction of DNA arrays or can be cloned for a large variety of functional analyses. These tools allow a replacement of single-gene analysis with a highly efficient whole-genome analysis.**

The genome sequencing projects have generated and will continue to generate enormous amounts of sequence data. The genomes of *Saccharomyces cerevisiae*, *Escherichia coli*, *Haemophilus influenzae* (1), *Mycoplasma genitalium* (2), and *Methanococcus jannaschii* (3) have been completely sequenced. Other model organisms have had substantial portions of their genomes sequenced as well, including the nematode *Caenorhabditis elegans* (4) and the small flowering plant *Arabidopsis thaliana* (5). This massive and increasing amount of sequence information allows the development of novel experimental approaches to identify gene function.

One standard use of genome sequence data is to attempt to identify the functions of predicted open reading frames (ORFs) within the genome by comparison to genes of known function. Such a comparative analysis of all ORFs to existing sequence data is fast, simple, and requires no experimentation and is therefore a reasonable first step. While finding sequence homologies/motifs is not a substitute for experimentation, noting the presence of sequence homology and/or sequence motifs can be a useful first step in finding interesting genes, in designing experiments and, in some cases, predicting function. However, this type of analysis is frequently uninformative. For example, over one-half of new ORFs in *S. cerevisiae* have no known function (6). If this is the case in a well studied organism such as yeast, the problem will be even worse in organisms that are less well studied or less manipulable. A large, experimentally determined gene function database would make homology/motif searches much more useful.

Experimental analysis must be performed to thoroughly understand the biological function of a gene product. Scaling up from classical "cottage industry" one-gene-oriented approaches to whole-genome analysis would be very expensive and laborious. It is clear that novel strategies are necessary to efficiently pursue the next phase of the genome projects—whole-genome experimental analysis to explore gene expression, gene product function, and other genome functions. Model organisms, such as *S. cerevisiae*, will be extremely important in the development of novel whole-genome analysis techniques and, subsequently, in improving our understanding of other more complex and less manipulable organisms.

The genome sequence can be systematically used as a tool to understand ORFs, gene product function, and other genome regions. Toward this end, a directed strategy has been developed for exploiting sequence information as a means of providing information about biological function (Fig. 1). Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons—they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay (7). As a pilot study, synthetic primers were made on the 96-well automated multiplex oligonucleotide synthesizer (AMOS) instrument (8) (Fig. 2). These oligonucleotides were used to amplify each ORF on yeast chromosome V. The current version of this instrument can synthesize three plates of 96 oligonucleotides each (25 bases) in an 8-hr day. The amplification of the entire set of PCR products was then analyzed by gel electrophoresis (Fig. 3). Successful amplification of the proper length product on the first attempt was 95%. This project demonstrates that one can go directly from sequence information to biological analysis in a truly automated, totally directed manner.

These amplicons can be incorporated directly in arrays or the amplicons can be cloned. If the amplicons are to be cloned, novel sequences can be incorporated at the 5′ end of the oligonucleotide to facilitate cloning. One potential problem with cloning PCR products is that the cloned amplicons may contain sequence alterations that diminish their utility. One option would be to resequence each individual amplicon. However, this is expensive, inefficient, and time consuming. A faster, more cost-effective, and more accurate approach is to apply comparative sequencing by denaturing HPLC (9). This method is capable of detecting a single base change in a 2-kb heteroduplex. Longer amplicons can be analyzed by use of appropriate restriction fragments. If any change is detected in a clone, an alternate clone of the same region can be analyzed. Modifying the system to allow high throughput analysis by denaturing HPLC is also relatively simple and straightforward.

If amplicons are used directly on arrays without cloning, it is important to note that, even if single PCR product bands are observed on gels, the PCR products will be contaminated with various amounts of other sequences. This contamination has

[illegible]

[illegible] Biochemistry, Beckman Center B400, Stanford University, Stanford CA 94305-5307. e-mail: gilbert@cmgm.stanford.edu

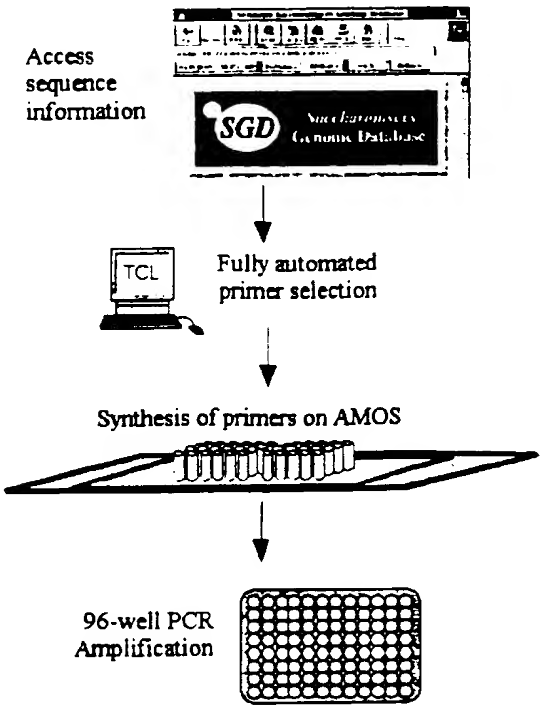

Access sequence information
SGD Saccharomyces Genome Database
TCL
Fully automated primer selection
Synthesis of primers on AMOS
96-well PCR Amplification

FIG. 1. Overview of systematic method for isolating individual genes. Sequence information is obtained automatically from sequence databases. The data are input into primer selection software specifically designed to target ORFs as designated by database annotations. The output file containing the primer information is directly read by a high-throughput oligonucleotide synthesizer, which makes the oligonucleotides in 96-well plates (AMOS, automated multiplex oligonucleotide synthesizer). The forward and reverse primers are synthesized in the same location on separate plates to facilitate the downstream handling of primers. The amplicons are generated by PCR in 96-well plates as well.

analysis. On the other hand, direct use of the amplicons is much less labor intensive and greatly decreases the occurrence of mistakes in clone identification, a ubiquitous problem associated with large clone set archiving and retrieving.

Any large-scale effort to capture each ORF within a genome must rely on automation if cost is to be minimized while efficiency is maximized. Toward that end, primers targeting ORFs were designed automatically using simple new scripts and existing primer selection software. These script-selected primer sequences were directly read by the high-throughput synthesizer and the forward and reverse primers were synthesized in separate plates in corresponding wells to facilitate automated pipetting and PCR amplifications. Each of the resulting PCR products, generated with minimum labor, contains a known, unique ORF.

Large-scale genome analysis projects are dependent on newly emerging technologies to make the studies practical and economically feasible. For example, the cost of the primers, a significant issue in the past, has been reduced dramatically to make feasible this and other projects that require tens of thousands of oligonucleotides. Other methods of high-throughput analysis are also vital to the success of functional analysis projects, such as microarraying and oligonucleotide chip methods (10–14).

Changes in attitude are also required. One of the major costs of commercial oligonucleotides is extensive quality control such that virtually 100% of the supplied oligonucleotides are successfully synthesized and work for their intended purpose. Considerable cost reduction can be obtained by simply decreasing the expected successful synthesis rate to 95–97%. One can then achieve faster and cheaper whole genome coverage by simply adding a single quality control at the end of the experiment and batching the failures for resynthesis.

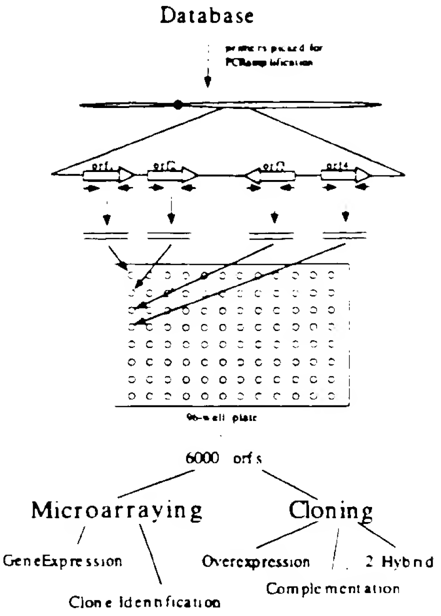


FIG. 2. Overall approach for using database of a genome to direct biological analysis. The synthesis of the 6,000 ORFs (orfs) for each gene of *S. cerevisiae* can be used in many applications utilizing both cloning and microarraying technology.

The directed nature of the amplicon approach is of clear advantage. The sequence of each ORF is analyzed automatically, and unique specific primers are made to target each ORF. Thus, there is relatively little time or labor involved—for example, no random cloning and subsequent screening is required because each product is known. In the test system, primers for 240 ORFs from chromosome V were systematically synthesized, beginning from the left arm and continuing through to the right arm. At no point was there any manual analysis of sequence information to generate the collection. In many ways, now that the sequence is known, there is no need for the researcher to examine it.

These amplicons can be arrayed and expression analysis can be done on all arrayed ORFs with a single hybridization (10). Those ORFs that display significant differential expression patterns under a given selection are easily identified without the laborious task of searching for and then sequencing a clone. Once scaled up, the procedure provides even greater returns on effort, because a single hybridization will ultimately provide a "snapshot" of the expression of all genes in the yeast genome. Thus, the limiting factor in whole genome analysis will not be the analysis process itself, but will instead be the ability of researchers to design and carry out experimental selections.

Current expression and genetic analysis technologies are geared toward the analysis of single genes and are ill suited to analyze numerous genes under many conditions. Additional difficulties with current technologies include: the effort and expense required to analyze expression and make mutants, the potential duplication of effort if done by different laboratories, and the possibility of conflicting results obtained from different laboratories. In contrast, whole genome analysis not only is more efficient, it also provides data of much higher quality; all genes are assayed and compared in parallel under exactly

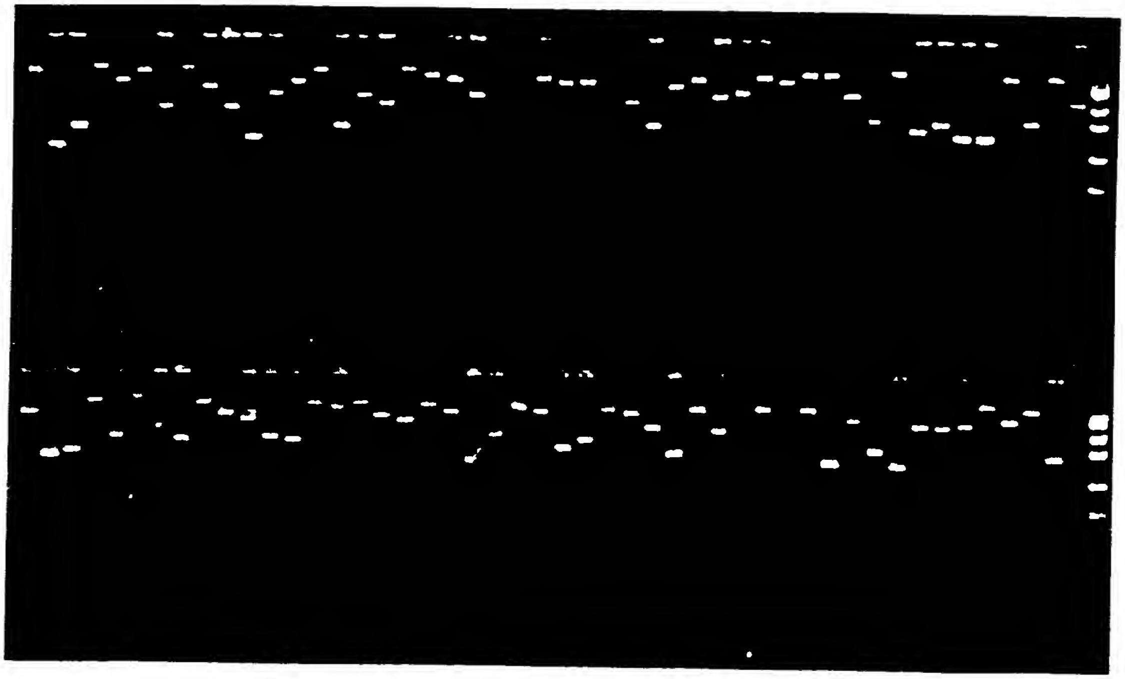


FIG. 3. Gel image of amplifications. Using the method described in Fig. 1, amplicons were generated for ORFs of *S. cerevisiae* chromosome V. One plate of 96 amplification reactions is shown.

the same conditions. In addition, amplicons have many applications beyond gene expression. For example, one recent approach is to incorporate a unique DNA sequence tag, synthesized as part of each gene specific primer, during amplification. The tags or molecular bar codes, when reintroduced into the organism as a gene deletion or as a gene clone, can be used much more efficiently than individual mutations or clones because pools of tagged mutants or transformants can be analyzed in parallel. This parallel analysis is possible because the tags are readily and quantitatively amplified even in complex mixtures of tags (13).

These ORF genome arrays and oligonucleotide tagged libraries can be used for many applications. Any conventional selection applied to a library that gives discrete or multiple products can use these technologies for a simple direct readout. These include screens and selections for mutant complementation, overexpression suppression (15, 16), second-site suppressors, synthetic lethality, drug target overexpression (17), two-hybrid screens (18), genome mismatch scanning (19) or recombination mapping.

The genome projects have provided researchers with a vast amount of information. These data must be used efficiently and systematically to gain a truly comprehensive understanding of gene function and, more broadly, of the entire genome which can then be applied to other organisms. Such global approaches are essential if we are to gain an understanding of the living cell. This understanding should come from the viewpoint of the integration of complex regulatory networks, the individual roles and interactions of thousands of functional gene products, and the effect of environmental changes on both gene regulatory networks and the roles of all gene products. The time has come to switch from the analysis of

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* **269**, 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* **270**, 397–403.
3. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al.* (1996) *Science* **273**, 1058–1073.
4. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. & Waterston, R. (1992) *Nature (London)* **356**, 37–41.
5. Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., *et al.* (1994) *Plant Physiol.* **106**, 1241–1255.
6. Oliver, S. (1996) *Nature (London)* **379**, 597–600.
7. Lashkari, D. A. (1996) Ph.D. dissertation (Stanford Univ., Stanford, CA).
8. Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W. & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 7912–7915.
9. Oefner, P. J. & Underhill, P. A. (1995) *Am. J. Hum. Genet.* **57**, A266.
10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270**, 467–470.
11. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991) *Science* **251**, 767–773.
12. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. & Fodor, S. P. (1996) *Science* **274**, 610–614.
13. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. (1996) *Nat. Genet.* **14**, 450–456.
14. Smith, V., Chou, K., Lashkari, D., Botstein, D. & Brown, P. O. (1996) *Science* **274**, 2069–2074.
15. Magdolen, V., Drubin, D. G., Mages, G. & Bandlow, W. (1993) *FEBS Lett.* **316**, 41–47.
16. Ramer, S. W., Elledge, S. J. & Davis, R. W. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 11589–11593.

IN PERSPECTIVE
Claudio J. Conti, Editor

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,[1] Michael Bittner,[2] Jeffrey Trent,[2] J. Carl Barrett,[1] and Cynthia A. Afshari[1]

[1] *Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina*
[2] *Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland*

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog. 24:153–159, 1999.* © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae,* was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cervisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNAse protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10–12] are possible solutions

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3′ regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

studies

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease–related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cervisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnostics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only 4n cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)+ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and BRCA1 [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

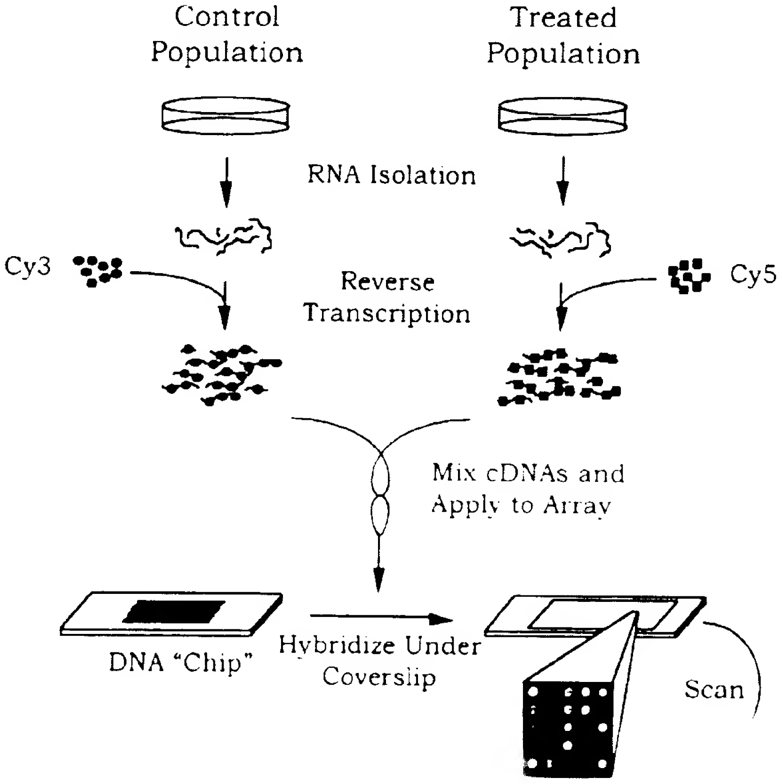
Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on relative fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.



Preparation and hybridization to cDNA microarrays. [illegible] although other sample types are amenable to this analysis

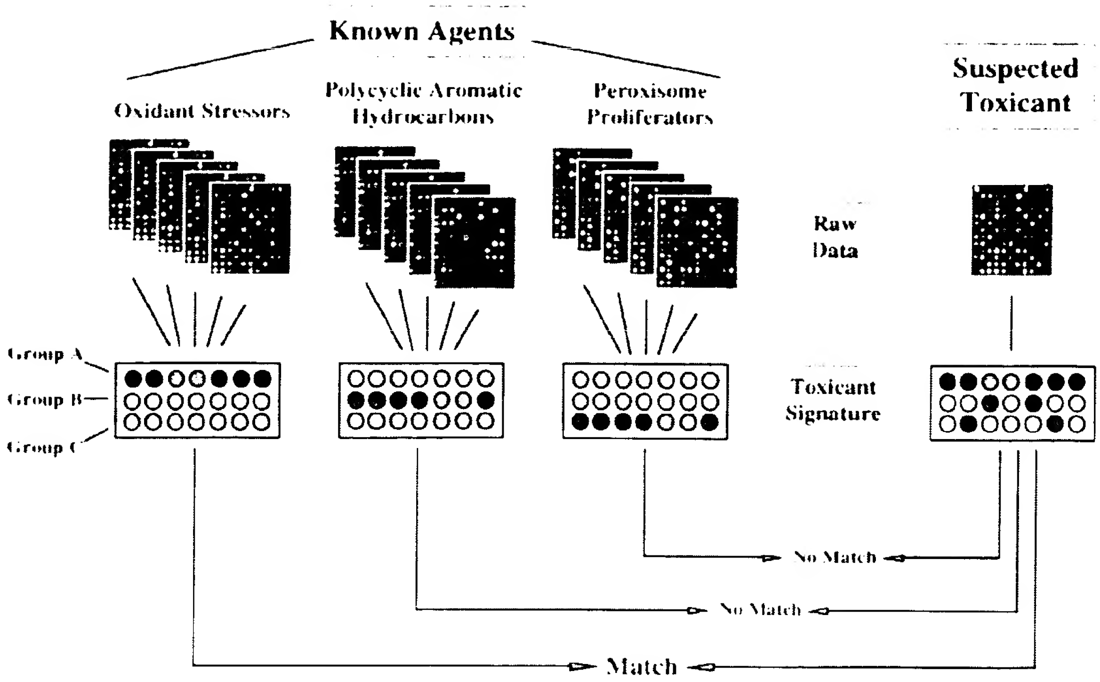


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive in vivo test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45].

Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24

REFERENCES

1 http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html
2 http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
3 Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496–512
4 Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. Science 1996;274:546, 563–567
5 http://www.perkin-elmer.com/press/prc5448.html
6 Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992;257:967–971
7 Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 1996;6:492–503
8 Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. Gene 1995;156:207–213
9 Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270:484–487
10 Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. Science 1995;270:467–470
11 DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–460
12 Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol 1997;15:1359–1367
13 Marshall A, Hodgson J. DNA chips: An array of possibilities. Nat Biotechnol 1998;16:27–31.
14 http://www.synteni.com
15 Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 1996;6:639–645
16 Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Biomedical Optics 1997;2:364–374
17 Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 1998;58:5009–5013
18 Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996;93:10614–10619

19 Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997;94:13057–13062.

20 Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 1997;94:2150–2155.

21 DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–686.

22 Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics 1996;37:29–40.

23 Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the E. coli genome. Genomics 1996;37:77–86.

24 Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. Biotechniques 1994;17:328–329, 332–336.

25 http://www.resgen.com/

26 http://www.genomesystems.com/

27 http://www.clontech.com/

28 Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society, 1992;203:34.

29 Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 1994;91:5022–5026.

30 Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251:767–773.

31 McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. Proc Natl Acad Sci USA 1996;93:13555–13560.

32 Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 1995;19:442–447.

33 Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675–1680.

34 http://www.mdyn.com/

35 Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. Genomics 1996;33:445–456.

36 Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610–614.

37 Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. Nat Genet 1998;18:155–158.

38 Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 1996;7:244–255.

39 Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. Nat Genet 1996;14:441–447.

40 Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat Med 1996;2:753–759.

41 Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–1082.

42 Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. Science 1998;281:1194–1197.

43 Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. Environ Health Perspect 1990;86:313–321.

44 Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. Nat Genet 1998;20:19–23.

45 http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html

46 Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 1996;382:722–725.

47 Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. J Natl Cancer Inst 1993;85:1159–1164.

48 http://www.niehs.nih.gov/envgenom/home.html

ELSEVIER

Toxicology Letters 112-113 (2000) 467-471

Toxicology Letters

www.elsevier.com/locate/toxlet

Expression profiling in toxicology — potentials and limitations

Sandra Steiner *, N. Leigh Anderson

Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA

Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers.

Keywords: Proteomics; Genomics; Toxicology

1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell

* Corresponding author. Tel.: +1-301-4245989; fax: +1-[illegible]

[illegible]

2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

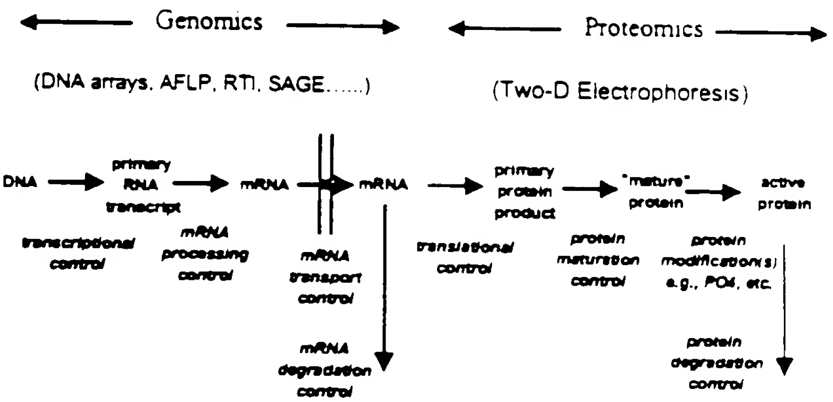


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

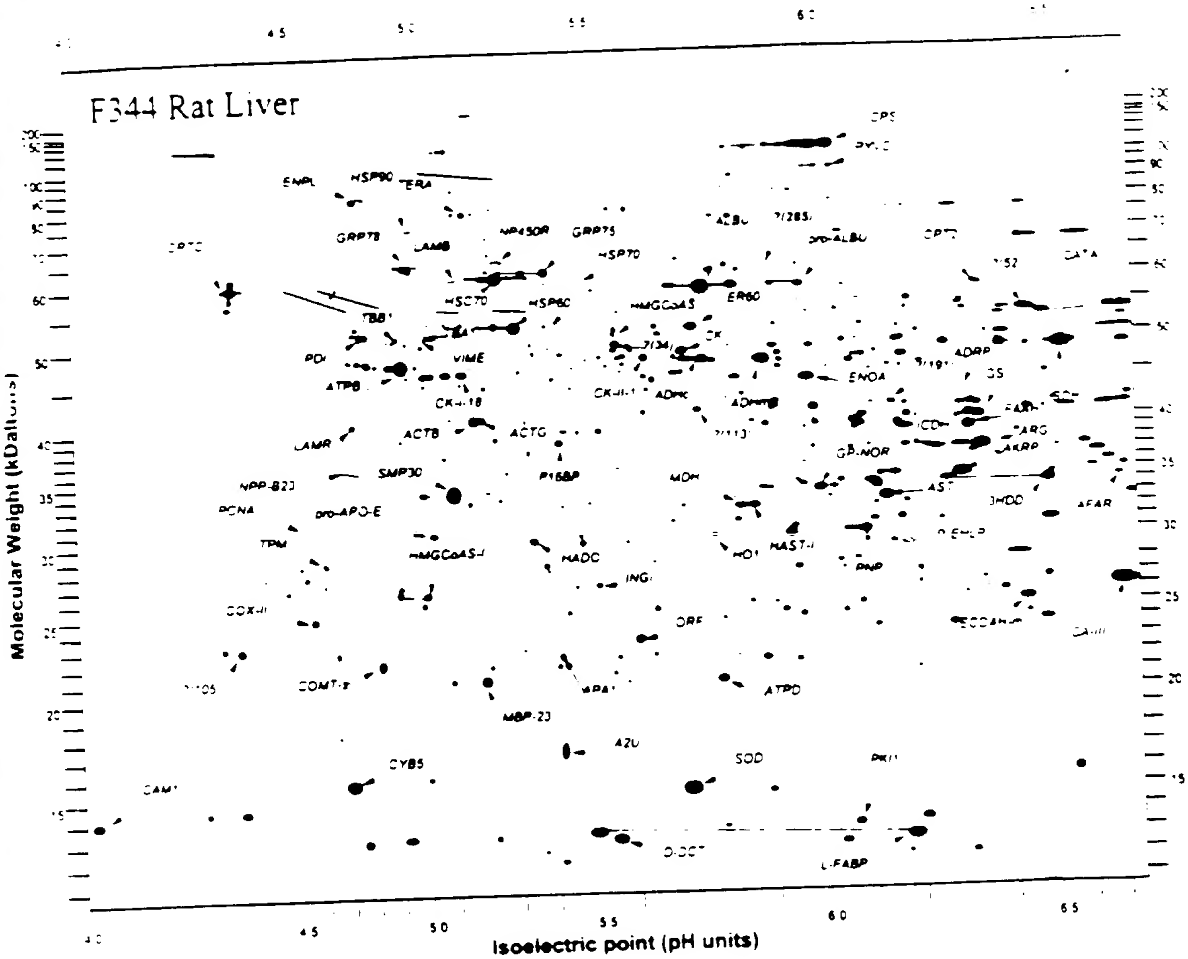


Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a

5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplification. Tissue biopsy samples typically yield good quality of both mRNA and proteins, however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug development, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trails.

References

Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. Electrophoresis 19, 1998–2003.

Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18, 533–537.

Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. Electrophoresis 12, 907–930.

Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. J. Cell. Biochem. Suppl. 22, 108–116.

Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. Toxicol. Appl. Pharmacol. 137, 75–89.

Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. Life Sci. 63, 2243–2250.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. Science 274, 610–614.

Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. Electrophoresis 19, 355–363.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. Science 251, 767–773.

Mann, M., Hojrup, P., Roepstorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22, 338–345.

Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. Electrophoresis 14, 157–161.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.

Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 6, 639–645.

Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Raymackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. Biochem. Biophys. Res. Commun. 218, 777–782.

Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. Biochem. Pharmacol. 51, 253–258.

Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. Curr. Biol. 6, 1543–1544.

Application of DNA Arrays to Toxicology

John C. Rockett and David J. Dix

Reproductive Toxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

DNA array technology makes it possible to rapidly genotype individuals or quantify the expression of thousands of genes on a single filter or glass slide, and holds enormous potential in toxicologic applications. This potential led to a U.S. Environmental Protection Agency-sponsored workshop titled "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. In addition to providing state-of-the-art information on the application of DNA or gene microarrays, the workshop catalyzed the formation of several collaborations, committees, and user's groups throughout the Research Triangle Park area and beyond. Potential application of microarrays to toxicologic research and risk assessment include genome-wide expression analyses to identify gene-expression networks and toxicant-specific signatures that can be used to define mode of action, for exposure assessment, and for environmental monitoring. Arrays may also prove useful for monitoring genetic variability and its relationship to toxicant susceptibility in human populations. *Key words:* DNA arrays, gene arrays, microarrays, toxicology.

Environ Health Perspect 107:681–685 (1999). [Online 6 July 1999]

http://ehpnet1.niehs.nih.gov/docs/1999/107p681-685rockett/abstract.html

Decoding the genetic blueprint is a dream that offers manifold returns in terms of understanding how organisms develop and function in an often hostile environment. With the rapid advances in molecular biology over the last 30 years, the dream has come a step closer to reality. Molecular biologists now have the ability to elucidate the composition of any genome. Indeed, almost 20 genomes have already been sequenced and more than 60 are currently under way. Foremost among these is the Human Genome Mapping Project. However, the genomes of a number of commonly used laboratory species are also under intensive investigation, including yeast, *Arabidopsis*, maize, rice, zebra fish, mouse, rat, and dog. It is widely expected that the completion of such programs will facilitate the development of many powerful new techniques and approaches to diagnosing and treating genetically and environmentally induced diseases which afflict mankind. However, the vast amount of data being generated by genome mapping will require new high-throughput technologies to investigate the function of the millions of new genes that are being reported. Among the most widely heralded of the new functional genomics technologies are DNA arrays, which represent perhaps the most anticipated new molecular biology technique since polymerase chain reaction (PCR).

Arrays enable the study of literally thousands of genes in a single experiment. The [illegible] ated to the technology (1). Despite this huge surge of interest, DNA arrays are still little used and largely unproven, as demonstrated by the high ratio of review and press articles to actual data papers. Even so, the potential they offer has driven venture capitalists into a frenzy of investment and many new companies are springing up to claim a share of this rapidly developing market.

The U.S. Environmental Protection Agency (EPA) is interested in applying DNA array technology to ongoing toxicologic studies. To learn more about the current state of the technology, the Reproductive Toxicology Division (RTD) of the National Health and Environmental Effects Research Laboratory (NHEERL; Research Triangle Park, NC) hosted a workshop on "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. The workshop was organized by David Dix, Robert Kavlock, and John Rockett of the RTD/NHEERL. Twenty-two intramural and extramural scientists from government, academia, and industry shared information, data, and opinions on the current and future applications for this exciting new technology. The workshop had more than 150 attendees, including researchers, students, and administrators from the EPA, the National Institute of Environmental Health Sciences (NIEHS), and a number of other establishments from Research Triangle Park and beyond. Presentations ranged from the technology behind array production through the sharing of actual experimental data and projections on the future importance and applications of arrays. The information contained [illegible]

Array Elements

In the context of molecular biology, the word "array" is normally used to refer to a series of DNA or protein elements firmly attached in a regular pattern to some kind of supportive medium. DNA array is often used interchangeably with gene array or microarray. Although not formally defined, microarray is generally used to describe the higher density arrays typically printed on glass chips. The DNA elements that make up DNA arrays can be oligonucleotides, partial gene sequences, or full-length cDNAs. Companies offering pre-made arrays that contain less than full-length clones normally use regions of the genes which are specific to that gene to prevent false positives arising through cross-hybridization. Sequence verification of cDNA clone identity is necessary because of errors in identifying specific clones from cDNA libraries and databases. Premade DNA arrays printed on membranes are currently or imminently available for human, mouse, and rat. In most cases they contain DNA sequences representing several thousand different sequence clusters or genes as delineated through the National Center for Biotechnology Information UniGene Project (2). Many of these different UniGene clusters (putative genes) are represented only by expressed sequence tags (ESTs).

Array Printing

Arrays are typically printed on one of two types of support matrix. Nylon membranes are used by most off-the-shelf array providers such as Clontech Laboratories, Inc. (Palo Alto, CA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). Microarrays such as those produced by Affymetrix, Inc. (Santa Clara, CA), Incyte Pharmaceuticals, Inc. (Palo Alto, CA), and many do-it-yourself (DIY) arraying groups use glass wafers or slides. Although standard microscope slides may be used, they must be preprepared to facilitate sticking of the DNA to the glass. Several different

Address correspondence to J. Rockett, Reproductive Toxicology Division (MD-72), National Health and Environmental Effects Research Laboratory, U.S. EPA, Research Triangle Park, NC 27711 USA. Telephone: (919) 541-2678. Fax: (919) 541-4017. E-mail: rockett.john@epa.gov

This document has been reviewed in accordance with EPA policy and approved for publication. Mention of companies, trade names, or products does not signify endorsement of such by the EPA.

Received 23 March [illegible] accepted 22 April 1999

coatings have been successfully used, including silane and lysine. The coating of slides can easily be carried out in the laboratory, but many prefer the convenience of precoated slides available from suppliers.

Once the support matrix has been prepared, the DNA elements can be applied by several methods. Affymetrix, Inc., has developed a unique photolithographic technology for attaching oligonucleotides to glass wafers. More commonly, DNA is applied by either noncontact or contact printing. Noncontact printers can use thermal, solenoid, or piezoelectric technology to spray aliquots of solution onto the support matrix and may be used to produce slide or membrane-based arrays. Cartesian Technologies, Inc. (Irvine, CA) has developed nQUAD technology for use in its PixSys printers. The system couples a syringe pump with the microsolenoid valve, a combination that provides rapid quantitative dispensing of nanoliter volumes (down to 4.2 nL) over a variable volume range. A different approach to noncontact printing uses a solid pin and ring combination (Genetic MicroSystems, Inc., Woburn, MA). This system (Figure 1) allows a broader range of sample, including cell suspensions and particulates, because the printing head cannot be blocked up in the same way as a spray nozzle. Fluid transfer is controlled in this system primarily by the pin dimensions and the force of deposition, although the nature of the support matrix and the sample will also affect transfer to some degree.

In contact printing, the pin head is dipped in the sample and then touched to the support matrix to deposit a small aliquot. Split pins were one of the first contact-printing devices to be reported and are the suggested format for DIY arrayers, as described by Brown (*3*). Split pins are small metal pins with a precise groove cut vertically in the middle of the pin tip. In this system, 1–48 split pins are positioned in the pin-head. The split pins work by simple capillary action, not unlike a fountain pen—when the pin heads are dipped in the sample, liquid is drawn into the pin groove. A small (fixed) volume is then deposited each time the split pins are gently touched to the support matrix. Sample (100–500 pL depending on a variety of parameters) can be deposited on multiple slides before refilling is required, and array densities of $> 2,500$ spots/cm^2 may be produced. The deposit volume depends on the split size, sample fluidity, and the speed of printing. Split pins are relatively simple to produce and can be made in-house if a suitable machine shop is available. Alternatively, they can be obtained directly from companies such as TeleChem International, Inc. (Sunnyvale, CA).

Irrespective of their source, printers should be run through a preprint sequence prior to producing the actual experimental arrays: the first 100 or so spots of a new run tend to be somewhat variable. Factors affecting spot reproducibility include slide treatment homogeneity, sample differences, and instrument errors. Other factors that come into play include clean ejection of the drop and clogging (nQUAD printing) and mechanical variations and long-term alteration in print-head surface of solid and split pins. However, with careful preparation it is possible to get a coefficient of variance for spot reproducibility below 10%.

One potential printing problem is sample carryover. Repeated washing, blotting, and drying (vacuum) of print pins between samples is normally effective at reducing sample carryover to negligible amounts. Printing should also be carried out in a controlled environment. Humidified chambers are available in which to place printers. These help prevent dust contamination and produce a uniform drying rate, which is important in determining spot size, quality, and reproducibility.

In summary, although several printing technologies are available, none are particularly outstanding and the bottom line is that they are still in a relatively early stage of evolution.

Array Hybridization

The hybridization protocol is, practically speaking, relatively straightforward and those with previous experience in blotting should have little difficulty. Array hybridizations are, in essence, reverse Southern/Northern blots—instead of applying a labeled probe to the target population of DNA/RNA, the labeled population is applied to the probe(s). With membrane-based arrays, the control and treated mRNA populations are normally converted to cDNA and labeled with isotope (e.g., ^{33}P) in the process. These labeled populations are then hybridized independently to parallel or serial arrays and the hybridization signal is detected with a phosphorimager. A less commonly used alternative to radioactive probes is enzymatic detection. The probe may be biotinylated, haptenylated, or have alkaline phosphatase/horseradish peroxidase attached. Hybridization is detected by enzymatic reaction yielding a color reaction (*4*). Differences in hybridization signals can be detected by eye or, more accurately, with the help of digital imaging and commercially available software. The labeling of the test populations for slide-based microarrays uses a slightly different approach. The probe typically consists of two samples of polyA$^+$ RNA (usually from a treated and a control population) that are converted to cDNA; in the process each is labeled with a different fluor. The independently labeled probes are then mixed together and hybridized to a single microarray slide and the resulting combined fluorescent signal is scanned. After normalization, it is possible to determine the ratio of fluorescent signals from a single hybridization of a slide-based microarray.

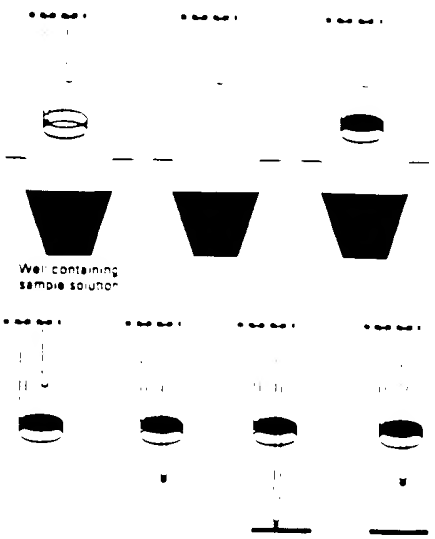


Figure 1. Genetic Microsystems (Woburn, MA) pin ring system for printing arrays. The pin ring combination consists of a circular open ring oriented parallel to the sample solution with a vertical pin centered over the ring. When the ring is dipped into a solution and lifted, it withdraws an aliquot of sample held by surface tension. To spot the sample, the pin is driven down through the ring and a portion of the solution is transferred to the bottom of the pin. The pin continues to move downward until the pendant drop of solution makes contact with the underlying surface. The pin is then lifted, and gravity and surface tension cause deposition of the spot onto the array. Figure from Flowers et al. (*14*), with permission from Genetic Microsystems.

cDNA derived from control and treated populations of RNA is most commonly hybridized to arrays, although subtractive hybridization or differential display reactions may also be used. Fluorophore- or radiolabeled nucleotides are directly incorporated into the cDNA in the process of converting RNA to cDNA. Alternatively, 5′ end-labeled primers may be used for cDNA synthesis. These are labeled with a fluorophore for direct visualization of the hybridized array. Alternatively, biotin or a hapten may be attached to the primer, in which case fluor-labeled streptavidin or antibody must be applied before a signal can be generated. The most commonly used fluorophores at present are cyanine (Cy)3 and Cy5 (Amersham Pharmacia Biotech AB, Uppsala, Sweden). However, the relative expense of these fluorescent conjugates has driven a search for cheaper alternatives. Fluorescein, rhodamine, and Texas red have all been used, and companies such as Molecular Probes, Inc. (Eugene, OR) are developing a series of labeled nucleotides with a wide range of excitation and emission spectra which may prove to function as well as the Cy dyes.

Analysis of DNA Microarrays

Membrane-based arrays are normally analyzed on film or with a phosphorimager, whereas chip-based arrays require more specialized scanning devices. These can be divided into three main groups: the charge-coupled device camera systems, the nonconfocal laser scanners, and the confocal laser scanners. The advantages and disadvantages of each system are listed in Table 1.

Table 1. Advantages and disadvantages of different microarray scanning systems

	CCD camera system	Nonconfocal laser scanner	Confocal laser scanner
Advantages	Few moving parts	Relatively simple optics	Small depth of focus reduces artifacts
	Fast scanning of bright samples	—	May have high light collection efficiency
Disadvantages	Less appropriate for dim samples	Low light collection efficiency	Small depth of focus requires scanning precision
	Optical scatter can limit performance	Background artifacts not rejected	
		Resolution typically low	

CCD, charge-coupled device
From Kawasaki (13)

Because a typical spot on a microarray can contain > 10^8 molecules, it is clear that a large variation in signal strength may occur. Current scanners cannot work across this many orders of magnitude (4 or 5 is more typical). However, the scanning parameters can normally be adjusted to collect more or less signal, such that two or three scans of the same array should permit the detection of rare and abundant genes.

When a microarray is scanned, the fluorescent images are captured by software normally included with the scanner. Several commercial suppliers provide additional software for quantifying array images, but the software tools are constantly evolving to meet the developing needs of researchers, and it is prudent to define one's own needs and clarify the exact capabilities of the software before its purchase. Issues that should be considered include the following:

- Can the software locate offset spots?
- Can it quantitate across irregular hybridization signals?
- Can the arrayed genes be programmed in for easy identification and location?
- Can the software connect via the Internet to databases containing further information on the gene(s) of interest?

One of the key issues raised at the workshop was the sensitivity of microarray technology. Experiments by General Scanning, Inc. (Watertown, MA), have shown that by using the Cy dyes and their scanner, signal can be detected down to levels of < 1 fluor molecule per square micrometer, which translates to detecting a rare message at approximately one copy per cell or less.

Array Applications

Although arrays are an emerging technology certain to undergo improvement and alteration, they have already been applied usefully to a number of model systems. Arrays are at their most powerful when they contain the [illegible] this reason, they have strong [illegible] *Caenorhabditis elegans* [illegible] The genomes of both of these species have been sequenced and, in the case of yeast, deposited onto arrays for examination of gene expression (6,7). With both of these species, it is relatively easy to perturb individual gene expression. Indeed, C. *elegans* knockouts can be made simply by soaking the worms in an antisense solution of the gene to be knocked out.

By a process of systematic gene disruption, it is now possible to examine the cause and effect relationships between different genes in these simple organisms. This kind of approach should help elucidate biochemical pathways and genetic control processes, deconvolute polygenic interactions, and define the architecture of the cellular network. A simple case study of how this can be achieved was presented by Butow [University of Texas Southwestern Medical Center, Dallas, TX (Figure 2)]. Although it is the phenotypic result of a single gene knockout that is being examined, the effect of such perturbation will almost always be polygenic. Polygenic interactions will become increasingly important as researchers begin to move away from single gene systems when examining the nature of toxicologic responses to external stimuli. This is especially important in toxicology because the phenotype produced by a given environmental insult is never the result of the action of a single gene; rather, it is a complex interaction of one or multiple cellular pathways. Phenomena such as quantitative trait (the continuous variation of phenotype), epistasis (the effect of alleles of one or more genes on the expression of other genes), and penetrance (proportion of individuals of a given genotype that display a particular phenotype) will become increasingly evident and important as toxicologists push toward the ultimate goal of matching the responses of individuals to different environmental stimuli.

Analysis of the transcriptome (the expression level of all the genes in a given cell population) was a use of arrays addressed by several speakers. Unfortunately, current gene nomen- [illegible] allocated [illegible] and there was [illegible] for standard [illegible] gene nomenclature. Nevertheless, once a transcriptome has been assembled it can then be transferred onto arrays and used to screen any chosen system. The EPA MicroArray Consortium (EPAMAC) is assembling tested transcriptomes for human, rat, and mouse. In a slightly different approach, Nuwaysir et al. (8) describes how the NIEHS assembled what is effectively a "toxicological transcriptome"—a library of human and mouse genes that have previously been proven or implicated in responses to toxicologic insults. Clontech Laboratories, Inc. (Palo Alto, CA), has begun a similar process by developing stress/toxicology filter arrays of rat, mouse, and human genes. Thus, rather than being tissue or cell specific, these stress/toxicology arrays can be used across a variety of model systems to look for alterations in the expression of toxicologically important genes and define the new field of toxicogenomics. The potential to identify toxicant families based on tissue- or cell-specific gene expression could revolutionize drug testing. These molecular signatures or fingerprints could not only point to the possible toxicity/carcinogenicity of newly discovered compounds (Figure 3), but also aid in elucidating their mechanism of action through identification of gene expression networks. By extension, such signatures could provide easily identifiable biomarkers to assess the degree, time, and nature of exposure.

DNA arrays are primarily a tool for examining differential gene expression in a given model. In this context they are referred to as closed systems because they lack the ability of other differential expression technologies, e.g., differential display and subtractive hybridization, to detect previously unknown genes not present on the array. This would appear to limit the power of DNA arrays to the imaginations and preconceptions of the researcher in selecting genes previously characterized and thought to be involved in the model system. However, the various genome sequencing projects have created a new category of sequence—the EST—that has partially molli- [illegible] may share [illegible] characterized genes have not been assigned specific genetic identity. By incorporating EST clones into an array, it is possible to monitor the expression of these unknown genes. This can enable the identification of previously uncharacterized genes that may have biologic significance in the model system. Filter arrays from Research Genetics and slide arrays from Incyte Pharmaceuticals both incorporate large numbers of ESTs from a variety of species.

A further use of microarrays is the identification of single nucleotide polymorphisms (SNPs). These genomic variations are abundant—they occur approximately every 1 kb or so—and are the basis of restriction fragment length polymorphism analysis used in forensic analysis. Affymetrix, Inc., designed chips that contain multiple repeats of the same gene sequence. Each position is present with all four possible bases. After the hybridization of the sample, the degree of hybridization to the different sequences can be measured and the exact sequence of the target gene deduced. SNPs are thought to be of vital importance in drug metabolism and toxicology. For example, single base differences in the regulatory region or active site of some genes can account for huge differences in the activity of that gene. Such SNPs are thought to explain why some people are able to metabolize certain xenobiotics better than others. Thus, arrays provide a further tool for the toxicologist investigating the nature of susceptible subpopulations and toxicologic response.

There are still many wrinkles to be ironed out before arrays become a standard tool for toxicologists. The main issues raised at the workshop by those with hands-on experience were the following:

- Expense: the cost of purchasing/contracting this technology is still too great for many individual laboratories.

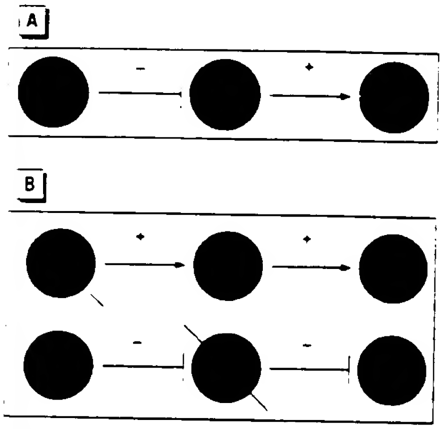


Figure 2. Potential effects of gene knockout within positively and negatively regulated gene expression networks. i_1 is limiting in wild type for expression of i_2. (A) A simple, two-component, linear regulatory network operating on gene i_2, where i_1 is a positive effector of i_2 and i_n is either a positive or negative effector of i_1. This network could be deduced by examining the consequence of (B) deleting i_n on the expression of i_1 and i_2, where the expression of i_2 would be decreased or increased depending on whether i_n was a positive or negative regulator. These and other connected components of even greater complexity could be revealed by genome-wide expression analysis. From Butow (15).

- Clones: the logistics of identifying, obtaining, and maintaining a set of nonredundant, noncontaminated, sequence-verified, species/cell/tissue/field-specific clones.
- Use of inbred strains: where whole-organism models are being used, the use of inbred strains is important to reduce the potentially confusing effects of the individual variation typically seen in outbred populations.
- Probe: the need for relatively large amounts of RNA, which limits the type of sample (e.g., biopsy) that can be used. Also, different RNA extraction methods can give different results.
- Specificity: the ability to discriminate accurately between closely related genes (e.g., the cytochrome p450 family) and splice variants.
- Quantitation: the quantitation of gene expression using gene arrays is still open to debate. One reason for this is the different incorporation of the labeling dyes. However, the main difficulty lies in knowing what to normalize against. One option is to include a large number of so-called housekeeping genes in the array. However, the expression of these genes often change depending on the tissue and the toxicant, so it is necessary to characterize the expression of these genes in the model system before utilizing them. This is clearly not a viable option when screening multiple new compounds. A second option is to include on the array genes from a nonrelated species (e.g., a plant gene on an animal array) and to spike the probe with synthetic RNA(s) complementary to the gene(s).
- Reproducibility: this is sometimes questionable, and a figure of approximately two or three repeats was used as the minimum number required to confirm initial findings. Again, however, most people advocated the use of Northern blots or reverse transcriptase PCR to confirm findings.
- Sensitivity: concerns were voiced about the number of target molecules that must be present in a sample for them to be detected on the array.
- Efficiency: reproducible identification of 1.5- to 2-fold differences in expression was reported, although the number of genes that undergo this level of change and remain undetected is open to debate. It is important that this level of detection be ultimately achieved because it is commonly perceived that some important transcription factors and their regulators respond at such low levels. In most cases, 3- to 5-fold was the minimum change that most were happy to accept.
- Bioinformatics: perhaps the greatest concern was how to accurately interpret the data with the greatest accuracy and efficiency. The biggest headache is trying to identify networks of gene expression that are common to different treatments or doses. The amount of data from a single experiment is huge. It may be that, in the future, several groups individually equipped with specialized software algorithms for studying their favorite genes or gene systems will be able to share the same hybridized chips. Thus, arrays could usher in a new perspective on collaboration and the sharing of data.

EPAMAC

Perhaps the main reason most scientists are unable to use array technology is the high cost involved, whether buying off-the-shelf membranes, using contract printing services, or

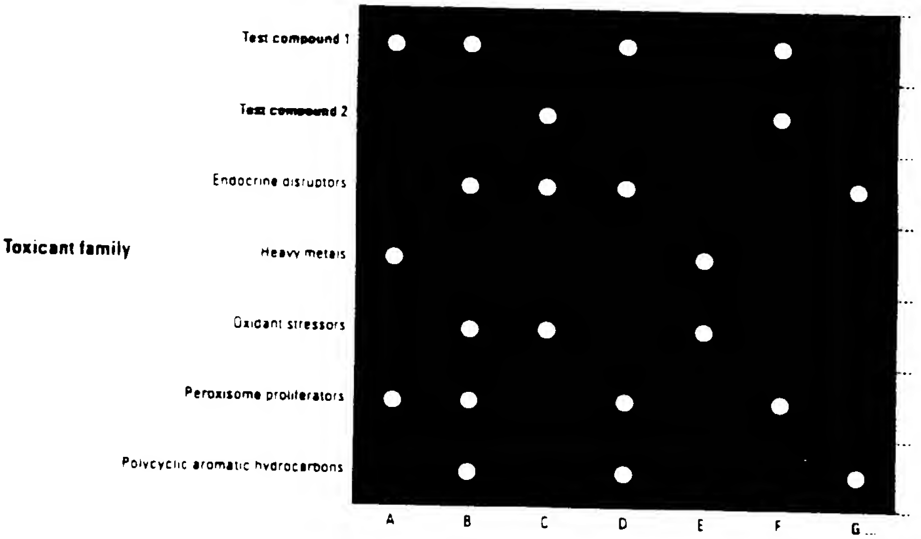


Figure 3. Gene expression profiles—also called fingerprints or signatures—of known toxicants or toxicant families may, in the future, be used to identify the potential toxicity of new drugs, etc. In this example, the genetic signature of test compound 1 is identical to that of known peroxisome proliferators, whereas that of test compound 2 does not match any known toxicant family. Based on these results, test compound 2 would be retained for further testing and test compound 1 would be eliminated.

producing chips in-house. In view of this, researchers at the RTD/NHEERL initiated the EPAMAC. This consortium brings together scientists from the EPA and a number of extramural labs with the aim of developing microarray capability through the sharing of resources and data. EPAMAC researchers are primarily interested in the developmental and toxicologic changes seen in testicular and breast tissue, and a portion of the workshop was set aside for EPAMAC members to share their ideas on how the experimental application of microarrays could facilitate their research. One of the central areas of interest to EPAMAC members is the effect of xenobiotics on male fertility and reproductive health. Of greatest concern is the effect of exposure during critical periods of development and germ cell differentiation (9), and how this may compromise sperm counts and quality following sexual maturation (10). As well as spermatogenic tissue, there is also interest in how residual mRNA found in mature sperm (11) could be used as an indicator of previous xenobiotic effects (it is easier to obtain a semen sample than a testicular biopsy). Arrays will be used to examine and compare the effect of exposure to heat and chemicals in testicular and epididymal gene expression profiles, with the aim of establishing relationships/associations between changes in developmental landmarks and the effects on sperm count and quality. Cluster, pattern, and other analysis of such data should help identify hidden relationships between genes that may reveal potential mechanisms of action and uncover roles for genes with unknown functions.

Summary

The full impact of DNA arrays may not be seen for several years, but the interest shown at this regional workshop indicates the high level of interest that they foster. Apart from educating and advertising the various technologies in this field, this workshop brought together a number of researchers from the Research Triangle Park area who are already using DNA arrays. The interest in sharing ideas and experiences led to the initiation of a Triangle array user's group.

Array technology is still in its infancy. This means that the hardware is still improving and there is no current consensus for standard procedures, quantitation, and interpretation. Consistency in spotting and scanning arrays is not yet optimized, and this is one of the most critical requirements of any experiment. In addition, one of the dark regions of array technology—strife in the courts over who owns what portions of it—has further muddled the future and is a potential barrier toward the development of consensus procedures.

Perhaps the greatest hurdle for the application of arrays is the actual interpretation of data. No specialists in bioinformatics attended the workshop, largely because they are rare and because as yet no one seems clear on the best method of approaching data analysis and interpretation. Cross-referencing results from multiple experiments (time, dose, repeats, different animals, different species) to identify commonly expressed genes is a great challenge. In most cases, we are still a long way from understanding how the expression of gene *X* is related to the expression of gene *Y*, and ordering gene expression to delineate causal relationships.

To the ordinary scientist in the typical laboratory, however, the most immediate problem is a lack of affordable instrumentation. One can purchase premade membranes at relatively affordable prices. Although these may be useful in identifying individual genes to pursue in more detail using other methods, the numbers that would be required for even a small routine toxicology experiment prohibit this as a truly viable approach. For the toxicologist, there is a need to carry out multiple experiments—dose responses, time curves, multiple animals, and repeats. Glass-based DNA arrays are most attractive in this context because they can be prepared in large batches from the same DNA source and accommodate control and treated samples on the same chip. Another problem with current off-the-shelf arrays is that they often do not contain one or more of the particular genes a group is interested in. One alternative is to obtain and/or produce a set of custom clones and have contract printing of membranes or slides carried out by a company such as Genomic Solutions, Inc. (Ann Arbor, MI). This approach is less expensive than buying or setting up one's own entire system, although at some point it might make economic sense to print one's own arrays.

Finally, DNA arrays are currently a team effort. They are a technology that uses a wide range of skills including engineering, statistics, molecular biology, chemistry, and bioinformatics. Because most individuals are skilled in only one or perhaps two of these areas, it appears that success with arrays may be best expected by teams of collaborators consisting of individuals having each of these skills.

Those considering array applications may be amused or goaded on by the following quote from *Fortune* magazine (12):

> Microprocessors have reshaped our economy, spawned vast fortunes and changed the way we live. Gene chips could be even bigger.

Although this comment may have been designed to excite the imagination rather than accurately reflect the truth, it is fair to say that the age of functional genomics is upon us. DNA arrays look set to be an important tool in this new age of biotechnology and will likely contribute answers to some of toxicology's most fundamental questions.

Speakers

Cindy Afshan
NIEHS
[illegible] Birnba[illegible]
[illegible]
[illegible]thwestern Medical Center
Alex Chenchik
Clontech Laboratories, Inc.
David Dix
U.S. EPA

Abdel Elkahloun
Research Genetics, Inc.
[illegible] Fenton
[illegible]man Ha[illegible]
[illegible] Burba
Paradigm Genetics
Bob Kavlock
U.S. EPA
Ernie Kawasaki
General Scanning, Inc.

Steve Krawetz
Wayne State University
[illegible] Mace
[illegible] Mordecai
Kevin Morgan
Glaxo Wellcome
Elaine Poplin
Research Genetics, Inc.
Don Rose
Cartesian Technologies, Inc.

Jim Samet
U.S. EPA
[illegible] Ward
[illegible] Weir
[illegible]
University of California
at Davis
Tim Zacharewski
Michigan State University

References and Notes

1. The chipping forecast. Nat Genet 21(Suppl 1):3–60 (1999).
2. National Center for Biotechnology Information. The Unigene System. Available: www.ncbi.nlm.nih.gov/Schuler/UniGene [cited 22 March 1999].
3. Brown PO. The Brown Lab. Available: http://cmgm.Stanford.edu/pbrown [cited 22 March 1999].
4. Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Chang F, et al. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics 51:313–324 (1998).
5. Ward S. DNA Microarray Technology to Identify Genes Controlling Spermatogenesis. Available: www.mcb.arizona.edu/wardlab/microarray.html [cited 22 March 1999].
6. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4:1293–1301 (1998).
7. Brown PO. The Full Yeast Genome on a Chip. Available: http://cmgm.stanford.edu/pbrown/yeastchip.html [cited 22 March 1999].
8. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshan CA. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 24(3):153–159 (1999).
9. Hecht NB. Molecular mechanisms of male germ cell differentiation. Bioessays 20:555–561 (1998).
10. Zacharewski TR. Timothy R. Zacharewski. Available: www.bch.msu.edu/faculty/zachar.htm [cited 22 March 1999].
11. Kramer JA, Krawetz SA. RNA in spermatozoa: implications for the alternative haploid genome. Mol Hum Reprod 3:473–478 (1997).
12. Stipp D. Gene chip breakthrough. Fortune, March [illegible]
13. [illegible] unpublished data.
14. Rogers YH, [illegible], Honkanen P, Montagu J, Rose SD. Development and Performance of a Novel Microarraying System Based on Surface Tension Forces. Available: http://www.geneticmicro.com/resources/html/coldspring.html [cited 22 March 1999].
15. Butow R (University of Texas Medical Center, Dallas, TX), unpublished data.

Subject: RE: [Fwd: Toxicology Chip]
Date: Mon. 3 Jul 2000 08:09:45 -0400
From: "Afshari.Cynthia" <afshari@niehs.nih.gov>
To: "'Diana Hamlet-Cox'" <dianahc@incyte.com>

You can see the list of clones that we have on our 12K chip at
http://manuel.niehs.nih.gov/maps/guest/clonesrch.cfm
We selected a subset of genes (2000K) that we believed critical to tox response and basic cellular processes and added a set of clones and ESTs to this. We have included a set of control genes (80+) that were selected by the NHGRI because they did not change across a large set of array experiments. However, we have found that some of these genes change significantly after tox treatments and are in the process of looking at the variation of each of these 80+ genes across our experiments.
Our chips are constantly changing and being updated and we hope that our data will lead us to what the toxchip should really be.
I hope this answers your question.
Cindy Afshari

> ----------
> *From:* *Diana Hamlet-Cox*
> *Sent:* *Monday, June 26, 2000 8:52 PM*
> *To:* *afshari@niehs.nih.gov*
> *Subject:* *[Fwd: Toxicology Chip]*
>
> *Dear Dr. Afshari,*
>
> *Since I have not yet had a response from Bill Grigg, perhaps he was not the right person to contact.*
>
> *Can you help me in this matter? I don't need to know the sequences, necessarily, but I would like very much to know what types of sequences are being used, e.g., GPCRs (more specific?), ion channels, etc.*
>
> *Diana Hamlet-Cox*
>
> *-------- Original Message --------*
> *Subject: Toxicology Chip*
> *Date: Mon, 19 Jun 2000 18:31:48 -0700*
> *From: Diana Hamlet-Cox <dianahc@incyte.com>*
> *Organization: Incyte Pharmaceuticals*
> *To: grigg@niehs.nih.gov*
>
> *Dear Colleague:*
>
> *I am doing literature research on the use of expressed genes as pharmacotoxicology markers, and found the Press Release dated February 29, 2000 regarding the work of the NIEHS in this area. I would like to know if there is a resource I can access (or you could provide?) that would give me a list of the 12,000 genes that are on your Human ToxChip Microarray. In particular, I am interested in the criteria used to select sequences for the ToxChip, including any control sequences included in the microarray.*
>

> --
>
> ======================

```
> This email message is for the sole use of the intended recipient(s) and
> may contain confidential and privileged information subject to
> attorney-client privilege. Any unauthorized review, use, disclosure or
> distribution is prohibited.  If you are not the intended recipient,
> please contact the sender by reply email and destroy all copies of the
> original message.
> ========================
>
>
```

Article No. jmbi.1999.2661 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) 288, 147–164

JMB

The Relationship between Protein Structure and Function: a Comprehensive Survey with Application to the Yeast Genome

Hedi Hegyi and Mark Gerstein*

Department of Molecular
Biophysics & Biochemistry
Yale University, 266 Whitney
Avenue, PO Box 208114
New Haven, CT, 06520 USA

**Corresponding author*

For most proteins in the genome databases, function is predicted *via* sequence comparison. In spite of the popularity of this approach, the extent to which it can be reliably applied is unknown. We address this issue by systematically investigating the relationship between protein function and structure. We focus initially on enzymes functionally classified by the Enzyme Commission (EC) and relate these to by structurally classified domains the SCOP database. We find that the major SCOP fold classes have different propensities to carry out certain broad categories of functions. For instance, alpha/beta folds are disproportionately associated with enzymes, especially transferases and hydrolases, and all-alpha and small folds with non-enzymes, while alpha + beta folds have an equal tendency either way. These observations for the database overall are largely true for specific genomes. We focus, in particular, on yeast, analyzing it with many classifications in addition to SCOP and EC (i.e. COGs, CATH, MIPS), and find clear tendencies for fold-function association, across a broad spectrum of functions. Analysis with the COGs scheme also suggests that the functions of the most ancient proteins are more evenly distributed among different structural classes than those of more modern ones. For the database overall, we identify the most versatile functions, i.e. those that are associated with the most folds, and the most versatile folds, associated with the most functions. The two most versatile enzymatic functions (hydro-lyases and *O*-glycosyl glucosidases) are associated with seven folds each. The five most versatile folds (TIM-barrel, Rossmann, ferredoxin, alpha-beta hydrolase, and P-loop NTP hydrolase) are all mixed alpha-beta structures. They stand out as generic scaffolds, accommodating from six to as many as 16 functions (for the exceptional TIM-barrel). At the conclusion of our analysis we are able to construct a graph giving the chance that a functional annotation can be reliably transferred at different degrees of sequence and structural similarity. Supplemental information is available from http://bioinfo.mbb.yale.edu/genome/foldfunc.

Keywords: structure-function; fold classification; structural convergence; functional divergence; yeast genomics

Introduction

The problem of determining function from sequence

[illegible] genome analysis is to deter

ducts in a genome. However, the function of only a minor fraction of proteins has been studied experimentally, and, typically, prediction of function is based on sequence similarity with proteins of known function. That is, functional annotation is transferred based on similarity. Unfortunately,

(1998), in particular, has noted that transferring of incorrect functional information threatens to

[illegible]
E-mail address of the corresponding author: Mark.Gerstein@yale.edu

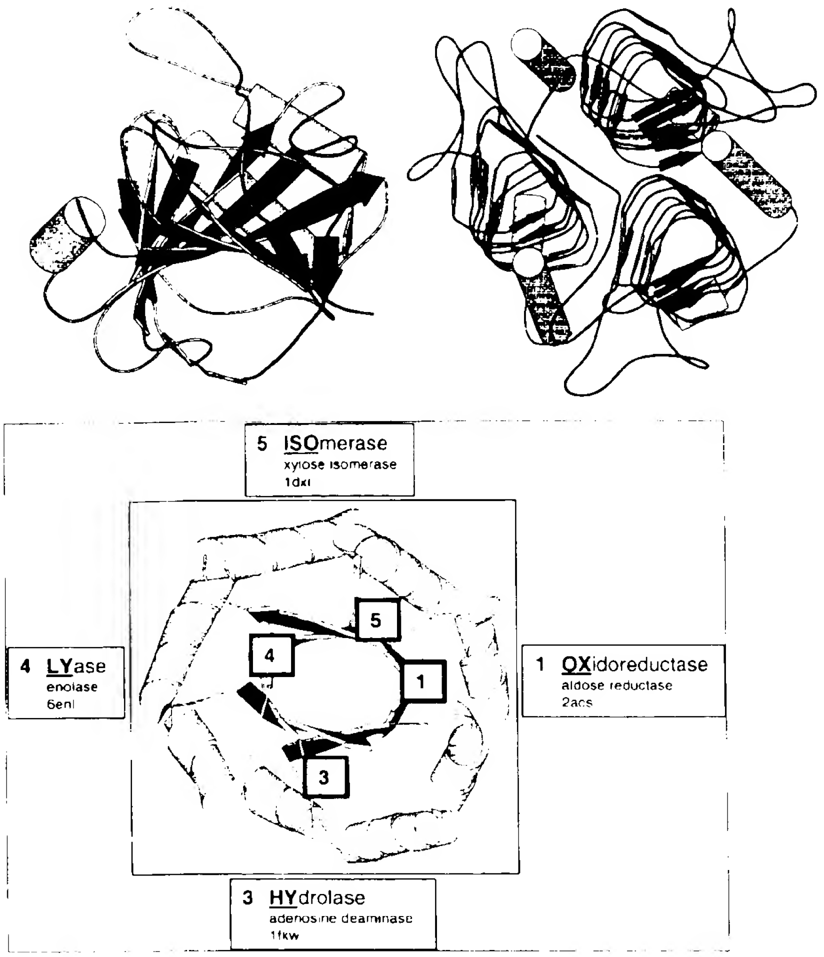


Figure 1. Specific example of convergent and divergent evolution. Top, an example of convergent evolution, showing structures of two carbonic anhydrases with the same enzymatic function (EC number 4.2.1.1), but with different folds. The Figure was drawn with Molscript (Kraulis, 1991) from 1THJ (left-handed beta helix) and 1DMX (flat beta sheet). Bottom, an example of possible divergent evolution, the TIM-barrel. This fold functions as a generic scaffold catalyzing 15 different enzymatic functions. A schematic Figure of the TIM-barrel fold is shown with numbers in boxes indicating the different location of the active site in four proteins that have this fold. These four proteins, xylose isomerase, aldose reductase, enolase, and adenosine deaminase, carry out very different enzymatic functions, in four of the main EC classes (1.*.*, 3.*.*, 4.*.*, and 5.*.*). They have active sites at very different locations (identified by the boxed numbers in the barrel) yet they all share the same fold.

progressively corrupt genome databases through the problem of accumulating incorrect annotations and using them as a basis for further annotations, and so on.

It is known that sequence similarity does confer structural similarity. Moreover, there is a well-established quantified relationship between the extent of similarity in sequence and that in structure. First investigated by Chothia & Lesk (1986) the similarity between the structures of two proteins (in terms of RMS) appears to be a monotonic function of their sequence similarity. This fact is often exploited when two sequences are declared related, based on a database search by programs such as BLAST or FastA (Altschul *et al.*, 1997; Pearson, 1996). Often, the only common element in two distantly related protein sequences is their underlying structures, or folds.

Transitivity requires that the well-established relationship between sequence and structure, and the more indefinite one between sequence and function, imply an indefinite relationship between structure and function. Several recent papers have highlighted this, analyzing individual protein superfamilies with a single fold but diverse functions. Examples include the aldo-keto reductases, a large hydrolase superfamily, and the thiol protein esterases. The latter include the eye-lens and corneal crystallins, a remarkable example of functional divergence (Bork & Eisenberg, 1998; Bork *et al.*, 1994; Cooper *et al.*, 1993; Koonin & Tatusov, 1994; Seery *et al.*, 1998).

There are also many classic examples of the converse: the same function achieved by proteins with completely different folds. For instance, even though mammalian chymotrypsin and bacterial subtilisin have different folds, they both function as serine proteases and have the same Ser-Asp-His catalytic triad. Other examples include sugar kinases, anti-freeze glycoproteins, and lysyl-tRNA synthetases (Bork *et al.*, 1993; Chen *et al.*, 1997; Doolittle, 1994; Ibba *et al.*, 1997a,b).

Figure 1 shows well-known examples of each of these two basic situations: the same fold but different function (divergent evolution) and the same function but different fold (convergent evolution).

Protein classification systems

The rapid growth in the number of protein sequences and three-dimensional structures has made it practical and advantageous to classify proteins into families and more elaborate hierarchical systems. Proteins are grouped together on the basis of structural similarities in the FSSP (Holm & Sander, 1998), CATH (Orengo *et al.*, 1997), and SCOP databases (Murzin *et al.*, 1995). SCOP is based on the judgments of a human expert FSSP, on automatic methods, and CATH, on a mixture of both. Other databases collect proteins on the basis of sequence similarities to one another, e.g. PROSITE, SBASE, Pfam, BLOCKS, PRINTS and ProDom (Attwood *et al.*, 1998; Bairoch *et al.*, 1997; Corpet *et al.*, 1998; Fabian *et al.*, 1997; Henikoff *et al.*, 1998; Sonnhammer *et al.*, 1997). Several collections contain information about proteins from a functional point of view. Some of these focus on particular organisms, e.g. the MIPS functional catalogue and YPD for yeast (Mewes *et al.*, 1997; Hodges *et al.*, 1998) and EcoCyc and GenProtEC for *Escherichia coli* (Karp *et al.*, 1998; Riley, 1997). Others focus on particular functional aspects in multiple organisms, e.g. the WIT and KEGG databases, which focus on metabolism and pathways (Selkov *et al.*, 1997; Ogata *et al.*, 1999), the ENZYME database, which focuses obviously enough on enzymes (Bairoch, 1996), and the COGs system, which focuses on proteins conserved over phylogenetically distinct species (Tatusov *et al.*, 1997). The ENZYME database, in particular, contains all the enzyme reactions that have an Enzyme Commission (EC) number assigned in accordance with the International Nomenclature Committee and is cross-referenced with Swissprot (Bairoch, 1996; Bairoch & Apweiler, 1998; Barrett, 1997).

Our approach: systematic comparison of proteins classified by structure with those classified by function

One of the most valuable operations one can do to these individual classification systems is to cross-reference and cross-tabulate them, seeing how they overlap. We performed such an analysis here by systematically interrelating the SCOP, Swissprot and ENZYME databases (Bairoch, 1996; Bairoch & Apweiler, 1998; Murzin *et al.*, 1995). For yeast we also have used the MIPS yeast functional catalogue, CATH and COGs in our analysis. This enables us to investigate the relationship between protein function and structure in a comprehensive statistical fashion. In particular, we investigated the functional aspects of both divergent and convergent evolution, exploring cases where a structure gains a dramatically different biochemical function and finding instances of similar enzymatic functions performed by unrelated structures

We concentrated on single-domain Swissprot

which can have only one assigned fold [illegible] to establish a one-to-one relationship between structure and function

Recent related work

This work is following up on several recent reports on the relationship between protein structure and function. In particular, Martin *et al.* (1998) studied the relationship between enzyme function and the CATH fold classification. They concluded that functional class (expressed by top-level EC numbers) is not related to fold, since a few specific residues, not the whole fold, determine enzyme function. Russell (1998) also focused on specific side-chain patterns, arguing that these could be used to predict protein function. In a similar fashion, Russell *et al.* (1998) identified structurally similar "supersites" in superfolds. They estimated that the proportion of homologues with different binding sites, and therefore with different functions, is around 10%. In a novel approach, using machine learning techniques, des Jardins *et al.* (1997) predict purely from the sequence whether a given protein is an enzyme and also the enzyme class to which it belongs.

Our work is also motivated by recent work looking at whether or not organisms are characterized by unique protein folds (Frishman & Mewes, 1997; Gerstein, 1997, 1998a,b; Gerstein & Hegyi, 1998; Gerstein & Levitt, 1997). If function is closely associated with fold (in a one-to-one sense), one would think that when a new function arose in evolution, nature would have to invent a new fold. Conversely, if fold and function are only weakly coupled, one would expect to see a more uniform distribution of folds amongst organisms and a high incidence of convergent evolution. In fact, a recent study on microbial genome analysis claims that functional convergence is quite common (Koonin & Galperin, 1997). Another related paper systematically searched Swissprot for all such cases of what is termed "analogous" enzymes (Galperin *et al.*, 1998).

Our work is also motivated by the recent work on protein design and engineering which aims to rationally change a protein function, for instance, to engineer a reporter function into a binding protein (Hellinga, 1997, 1998; Marvin *et al.*, 1997).

Results

Overview of the 8937 single-domain matches

Our basic results were based on simple sequence comparisons between Swissprot and SCOP, the SCOP domain sequences being used as queries against Swissprot. We focused on "mono-functional" single-domain matches in Swissprot, i.e. those singe-domain proteins with only one anno[illegible] The detailed criteria used in the

are of known structure, and about one eighth are both. (More precisely, of the 69,113 analyzed pro teins in Swissprot, 19,995 are enzymes, 18,317 are structural homologues, and 8205 are both.) About half of the fraction of Swissprot that matched known structures were "single-domain" and about one-third of these were enzymes (8937 and 3359, respectively, of 18,317). We focus on these 8937 single-domain matches here. Notice how these numbers also show how the known structures are significantly biased towards enzymes: 45% (8205 out of 18,317) of all the structural homologues are enzymes *versus* 29% (19,995 out of 69,113) for all of Swissprot.

331 observed fold-function combinations

Figure 2 gives an overview of how the matches are distributed amongst specific functions and folds. The single-domain matches include 229 of the 361 folds in SCOP 1.35, and 91 of the 207 three-component enzyme categories in the ENZYME database (Bairoch, 1996). Each match combines a SCOP fold number on the structural side (columns in Figure 2) and a three-component EC category on the functional side (rows), with all the non-enzymatic functions grouped together into a single category with the artificial "EC number" of 0.0.0 (shown in the first row in Figure 2). This results in a table where each cell represents a potential fold-function combination. The table contains a maximum of 21,068 (= 229 × 92) possible fold-function combinations (and a minimum of 229 combinations, assuming only one function for every fold). We actually observe 331 of these combinations (1.6%, shown by the filled-in cells).

Overall, more than half of the functions are associated with at least two different folds, while less than half of the folds with enzymatic activity have at least two functions (51 out of 91 and 53 out of 128, respectively).

Summarizing the fold-function combinations by 42 broad structure-function classes

As listed in Table 1, folds can be subdivided in six broad fold classes (e.g. all-alpha, all-beta, alpha/beta, etc.). Likewise, functions can be broken into seven main classes, non-enzymes plus six enzyme classes, e.g. oxidoreductase, transferase, etc. This gives rise to 42 (6 × 7) structure-function classes. The way the 21,068 potential fold-function combinations are apportioned amongst the 42 classes is shown in Table 2A.

Table 2B shows the way the 331 observed combinations were actually distributed amongst the 42 classes. Comparing the number of possible combinations with that observed shows that the most densely populated region of the chart is the transferase, hydrolase and lyase functions in combi-

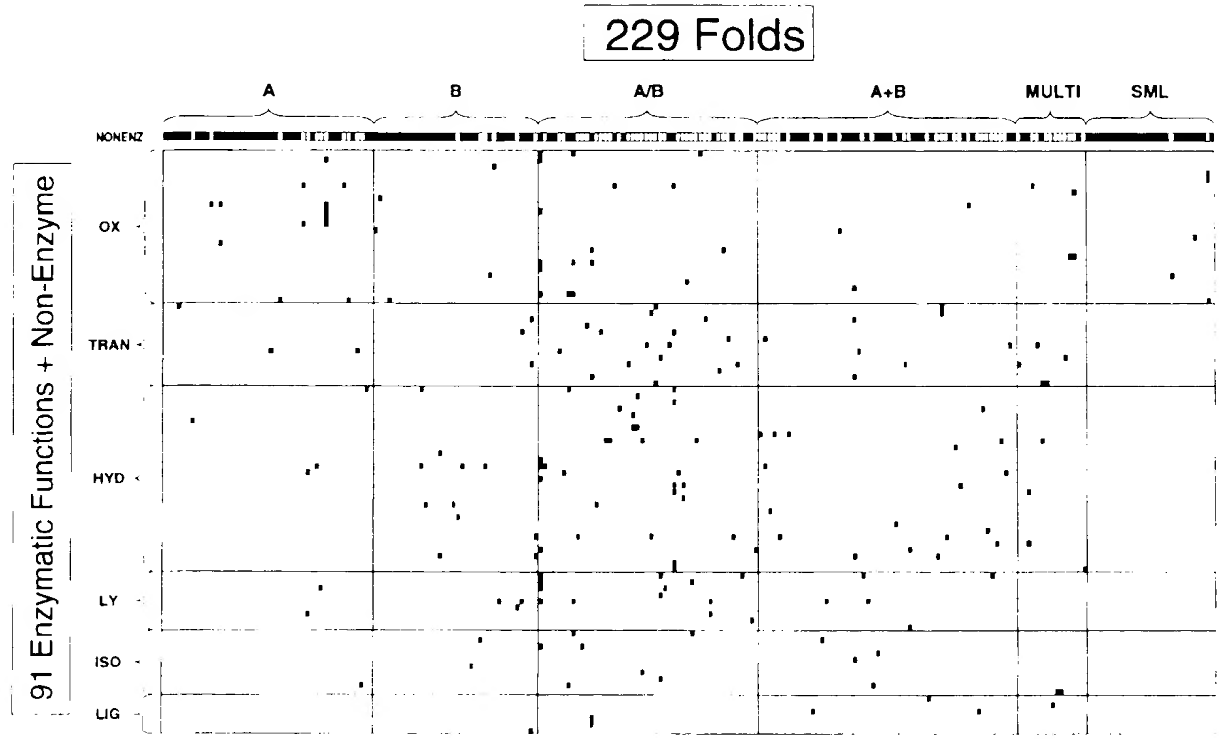


Figure 2. Overview of all the single-domain matches between proteins in Swissprot 35 and domains in SCOP 1.35. Sequences were compared with BLAST using the match criteria described in Materials and Methods. The matches are clustered into 92 functions (based on three-component EC numbers), which are arranged on each row, and 229 folds (based on SCOP fold numbers), which are arranged on each column. The first row indicates the matches with non-enzymes. There are, thus, 21,068 (= 92 × 229) possible combinations shown in the Figure. Only the 331 are actually observed. These are indicated by filled squares.

Table 1. Broad structural and functional categories

A *Functional categories in Swissprot 35*[a]

EC category	Category name	Abbreviation	Num. of functions in category
0.0.0	Non-enzymes	NONENZ	1
1.*.*	Oxidoreductases	OX	86
2.*.*	Transferases	TRAN	28
3.*.*	Hydrolases	HYD	53
4.*.*	Lyases	LY	15
5.*.*	Isomerases	ISO	16
6.*.*	Ligases	LIG	9
		Total:	208

B *Structural classes in SCOP 1.35*[b]

Fold class	Class name	Abbreviation	Num. of folds in class
1	All-alpha	A	81
2	All-beta	B	57
3	Alpha and beta	A/B	70
4	Alpha plus beta	A + B	91
5	Multi-domain	MULTI	19
6	Transmembrane	TM	9
7	Small proteins	SML	43
		Total:	361

[a] List of the functional (enzymatic) categories in Swissprot and the abbreviations used here. The values denote the number of three-component EC numbers in each category.

[b] List of the structural classes in SCOP studied here, and the abbreviations used for the classes. Values denote the number of folds in each class in SCOP 1.35. Class 6 is not used in the analysis.

nation with the alpha/beta fold class. This notion is in accordance with the general view that the most popular structures among enzymes fall into the alpha/beta class. In contrast, matches between small folds and enzymes are almost completely missing, except for five folds in the oxidoreductase category. There are also no all-alpha ligases and only one all-alpha isomerase.

Table 2C and D break down the 331 fold-function combinations in Table 2A into either just a number of folds or just a number of functions. That is, Table 2C lists the number of different folds associated with each of the 42 structure-function classes (corresponding to the non-zero columns in the relevant class in Figure 2), and Table 2D does the same thing for functions (non-zero rows in Figure 2). Comparing these tables back to the total number of combinations (Table 2A) reveals some interesting findings, keeping in mind that more functions than folds reveals probable divergence and that more folds than functions reveals probable convergence. For instance, the alpha/beta and alpha + beta fold classes contain similar numbers of folds, but the alpha/beta class has relatively more functions, perhaps reflecting a greater divergence. (Specifically, the alpha/beta class has 73 folds and 56 functions, while the alpha + beta class has 67 folds but only 35 functions.)

Table 2E shows the number of matching Swis- [illegible]

[illegible] 361 of the 1159 Swissprot sequences have matches with the immunoglobulin fold. These numbers are obviously, affected by the biases in Swissprot. On the other hand, if we compare the total matches in Table 2E with the total combinations in Table 2B it is clear that the numbers do not directly correlate. For instance, fewer hydrolases in Swissprot have matches with alpha/beta folds than with alpha + beta folds (295 *versus* 452), but the number of different combinations in the first case is 30, as opposed to only 18 in the second case. This suggests that our approach of counting combinations may not be as affected by the biases in the databanks as simply counting matches.

Table 2F and G give some rough indication of the statistical significance of the differences in the observed distribution of combinations. In Table 2F, using chi-squared statistics, we calculate for each individual structure class the chance that we could get the observed distribution of fold-function combinations over various functional classes if fold was not related to function. Then in Table 2G, we reverse the role of fold and function, and calculate the statistics for each functional class.

Enzyme *versus* non-enzyme folds

On the coarsest level, function can be divided amongst enzymes and non-enzymes. Of the 229 folds present in Figure 2, 93 are associated only with enzymes and 101 are associated only with [illegible]

Figure 3(a) shows a graphical view of the distribution of the different fold classes among these broadest functional categories. The distribution is far from uniform. The all-alpha fold class has 30 non-enzymatic representatives, but only 12 purely enzymatic folds and four folds with "mixed" (both types of) functions. This implies that a protein with an all-alpha fold has *a priori* roughly twice the chance of having a non-enzymatic function over an enzymatic one. The all-beta fold class has six enzymatic, 17 non-enzymatic and 13 mixed folds. In the alpha/beta class, 34 folds are associated only with enzymes and five folds only with non-enzymes, whereas in the alpha + beta class this ratio is more balanced, 28 "purely" enzymatic folds *versus* 22 purely non-enzymatic ones.

Table 2. Statistics over 42 structure-function classes

A. *Number of possible combinations between folds and functions in each of 42 classes (number of cells in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	46	36	48	56	15	28	229
OX	1104	864	1152	1344	360	672	5496
TRAN	598	468	624	728	195	364	2977
HYD	1334	1044	1392	1624	435	812	6641
LY	414	324	432	504	135	252	2061
ISO	460	360	480	560	150	280	2290
LIG	276	216	288	336	90	168	1374
Sum	4232	3312	4416	5152	1380	2576	21,068

B. *Number of observed combinations between folds and functions in each of 42 classes (number of filled cells in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	34	30	14	28	4	26	136
OX	13	5	17	3	4	5	47
TRAN	3	3	16	8	5		35
HYD	4	11	30	18	4		67
LY	2	3	13	5			23
ISO	1	2	7	4	2		16
LIG		1	2	3	1		7
Sum	57	55	99	69	20	31	331

C. *Number of folds in each of the 42 classes (columns with a filled cell in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	34	30	14	28	4	26	136
OX	7	5	9	3	3	3	30
TRAN	3	2	15	6	5		31
HYD	4	8	19	18	3		52
LY	2	3	8	5			18
ISO	1	2	7	4	2		16
LIG		1	1	3	1		6
Sum	51	51	73	67	18	29	289

D. *Number of functions in each of the 42 classes (rows with a filled cell in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	1	1	1	1	1	1	6
OX	8	5	9	3	3	5	33
TRAN	2	3	13	8	4		30
HYD	4	7	19	14	4		48
LY	2	2	7	3			14
ISO	1	2	5	4	1		13
LIG		1	2	2	1		6
Sum	18	21	56	35	14	6	150

E. *Total number of matching Swissprot sequences in each of the 42 fold-function classes*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	1940	1159	560	638	106	892	5295
OX	150	202	388	50	68	18	876
TRAN	65	14	363	116	174		732
HYD	116	394	295	452	92		1349
LY	40	47	168	104			359
ISO	2	54	122	22	2		202
LIG		5	26	69	24		124
Sum	2313	1875	1922	1451	466	910	8937

F. *How much does each of the fold classes deviate from the average distribution of functions?*

	χ^2	P
A	17.5	<0.01
B	5.2	<0.6
A/B	32.5	<0.00002
A + B	7.7	<0.3
MULTI	9.9	<0.2
SML	27.8	<0.0002

continued

Table 2—*Continued*

G. How much do each of the function classes deviate from the average distribution of folds?	χ^2	P
NONENZ	40.7	<0.0000002
OX	9.9	<0.08
TRAN	13.1	<0.03
HYD	17.3	<0.005
LY	10.2	<0.08
ISO	5.0	<0.5
LIG	4.3	<0.6

This Table shows various totals from Figure 2 distributed among the 42 structure-function classes, i.e. the seven functional categories in Table 1A multiplied by the six structural categories in Table 1B. Part A shows how many potential fold-function combinations there are in Figure 2 amongst each of the 42 classes. Part B shows how many of these 21,068 possible combinations are actually observed. Part C shows the total number of different folds (i.e. selected columns in Figure 1) in each class. Part D shows the total number of different functions (i.e. selected rows in Figure 2) in each class. Part E shows the total number of matching Swissprot proteins in the 42 classes. Note that to observe a fold-function combination one only needs the existence of a single match between a Swissprot protein and a SCOP domain. However, there can be many more. That is why the totals in this Table sum up to so much larger an amount than 331.

Here is an example of how to read parts A to E of the Table, focussing on the all-alpha, oxidoreductase region. Part A shows that there are 1104 cells, filled or unfilled, in this region, corresponding to possible combinations. Part B shows that 13 of these 1104 cells are filled, corresponding to observed all-alpha, oxidoreductase combinations. Part C shows that there are seven folds, corresponding to columns with filled cells in this region. Part D shows that there are eight functions, corresponding to rows with filled cells in this region. Finally, in part E we find that there are 150 Swissprot entries that have matches with a SCOP domain. They correspond to the 13 observed combinations in Part B.

Parts F and G give information on the statistical significance of the differences observed between the 42 structure-function classes. Part F gives the significance that the observed distribution of fold-function combinations in a given functional class is different than average (i.e. the null hypothesis that distribution of fold-function combinations is the same in each functional class). This is very similar to the derivation by Martin *et al.* (1998). A chi-squared statistic is computed for each of the seven functional classes in the conventional way: $\chi^2(f) = \Sigma_s (O_{sf} - E_{sf})^2/E_{sf}$, where for a given functional class f and structure class s, O_{sf} is the observed number of fold-function combinations and E_{sf} is the expected number. E_{sf} is simply computed from scaling the "sum" column and row in Part B of the Table: $E_{sf} = T_sT_f/T$, where T_s is the total number of combinations in a given structural class s (sum row), T_f is the total number of combinations in a given functional class f (sum column), and T is the total observed number of combinations, 331. Part G gives the statistical significance that the observed distribution of fold-function combinations in a given structural class is different than average. To compute this one simply sums over functions instead of structures: $\chi^2(s) = \Sigma_f(O_{sf} - E_{sf})^2/E_{sf}$. After each chi-squared statistic is reported, a rough probability or P-value is given. This gives the chance the observed distribution could be obtained randomly.

Restricting the comparison to individual genomes

Figure 3(a) applies to all of Swissprot. Figure 3(b) and (c) shows the functional distribution of folds taking into account the matches only in two specific genomes, yeast and *E. coli*. Only a fraction of each genome could be taken into consideration for various reasons (156 proteins in yeast, 244 proteins in *E. coli*), mostly due to the great number of enzymes having multiple domains in both yeast and *E. coli*. Chi-squared tests show that the fold distribution in yeast does not differ significantly from that in Swissprot and that the one in *E. coli* differs only slightly ($P < 0.25$ and $P < 0.02$, respectively). The main difference between Swissprot and *E. coli* is the larger fraction of alpha/beta enzymatic folds in the latter (34/93 *versus* 26/49). There are also somewhat more non-enzymatic all-alpha and small folds in Swissprot than in the two genomes. This is principally due to the greater prevalence of globins, myosins, cytochromes, toxins, and [illegible]

[illegible]

bution turns out to be similar to that of Swissprot (data not shown).

The yeast genome viewed from different classification schemes

In Figure 4 we focus on the yeast genome in more detail, trying to see the effect that different classification schemes have on our results. Although the total number of counts for our statistics decrease, in just using yeast relative to all of Swissprot, yeast provides a good reference frame to compare a number of classification schemes in as unbiased a fashion as possible. Also, yeast is one of the most comprehensively characterized organisms, and there are a number of functional classifications available exclusively for this organism.

In part Figure 4(a) we cross-tabulate the structure-function combinations in yeast using the SCOP and EC systems as we have done for all of Swissprot in Table 2B. The yeast distribution is fairly similar to that of Swissprot with the only major difference being somewhat more alpha/beta transferases and fewer alpha/beta hydrolases than expected. (A chi-squared test gives $P < \sim 0.05$ for the two distributions to differ. If either the transfer-

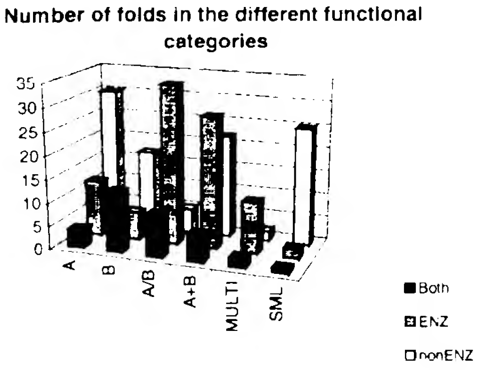
[illegible]

fication (Orengo *et al.*, 1997) instead of SCOP. For this Figure we mapped the SCOP classification of a yeast PDB match to its corresponding CATH classification and then cross-tabulated the structure-function combinations in the various classes. Essentially, this Figure shows the results reported by Martin *et al.* (1998) just for yeast.

In Figure 4(c) and (d), which show COGs *versus* SCOP cross-tabulations, we achieve the opposite of (b). We change the functional classifications scheme but keep SCOP for classifying structures. As was the case with the enzyme classification, but perhaps even more so, using COGs to classify function shows clearly that certain fold classes are associated with certain functions and *vice versa*. Most notably, whereas the functions associated with metabolism, which are mostly enzymes, are preferentially associated with the alpha/beta fold class, those associated with cellular processes (e.g. secretion) and information processing (e.g. transcription), show no such preference. They, in fact, show a marked preference for all-alpha structure. Small proteins are absent from most of the COGs classes, except one part of information processing and two in cellular processes.

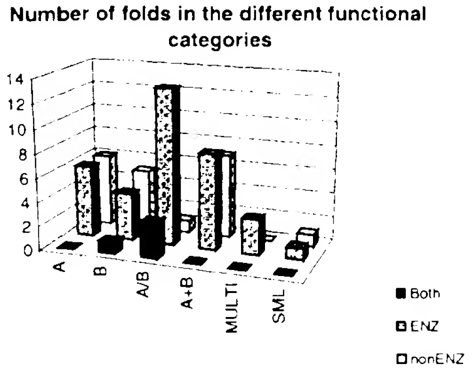
The COGs system classifies functions for those proteins that have clear orthologues in different species. Thus, conclusions based on using yeast COGs should be readily applicable to other genomes. This point is highlighted in Figure 4(d), which shows a COGs *versus* SCOP classification for only the 110 COGs that are conserved across all the analyzed genomes (eight) and all three kingdoms. Thus, this sub-figure would appear *exactly* the same for *E. coli*, *Methanococcos jannaschii* or a number of other genomes. It clearly shows how much more common the information processing proteins are among the most conserved and ancient proteins. Moreover, note how these most ancient proteins appear to have less of a preference for a particular structural class than the "more modern" metabolic ones. This suggests that large-scale duplication of alpha/beta folds for use in metabolism is what gave rise to stronger fold-function association in Figure 3(c).

A. All of Swissprot



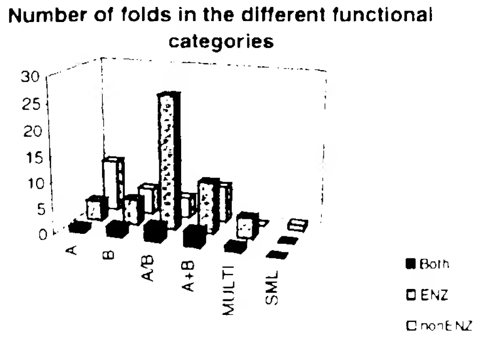
	A	B	A/B	A+B	MULTI	SML	TOTAL
Both	4	13	9	6	2	1	35
ENZ	12	6	34	28	11	2	93
nonENZ	30	17	5	22	2	25	101

B. Yeast



	A	B	A/B	A+B	MULTI	SML	TOTAL
Both	0	1	3	0	0	0	4
ENZ	6	4	13	8	3	1	35
nonENZ	6	5	1	7	0	1	20

C. E. coli



	A	B	A/B	A+B	MULTI	SML	TOTAL
Both	1	2	3	3	1	0	10
ENZ	4	5	26	10	4	0	49
nonENZ	10	5	4	7	0	1	27

Figure 3. Chart with breakdown among structure-function classes in two genomes. Charts and Tables showing the number of folds in each fold class associated with only enzymatic (ENZ), only non-enzymatic (nonENZ), and both enzymatic and non-enzymatic functions (Both). The results are shown for (a) all of Swissprot, (b) for just the yeast genome, and (c) for just the *E. coli* genome. The results for individual domains in a minimum set of SCOP domains also support these tendencies (data not shown). The numbers in (b) are not based on the PSI-blast protocol used for Figure 4. Rather they are found just as "subsets" of the overall Swissprot results to make them readily comparable with the rest of the paper. Because of this the numbers in this Figure will not match exactly those in Figure 4, the difference having to do with the greater number of fold-function combinations found by PSI-blast as compared to WU-blast

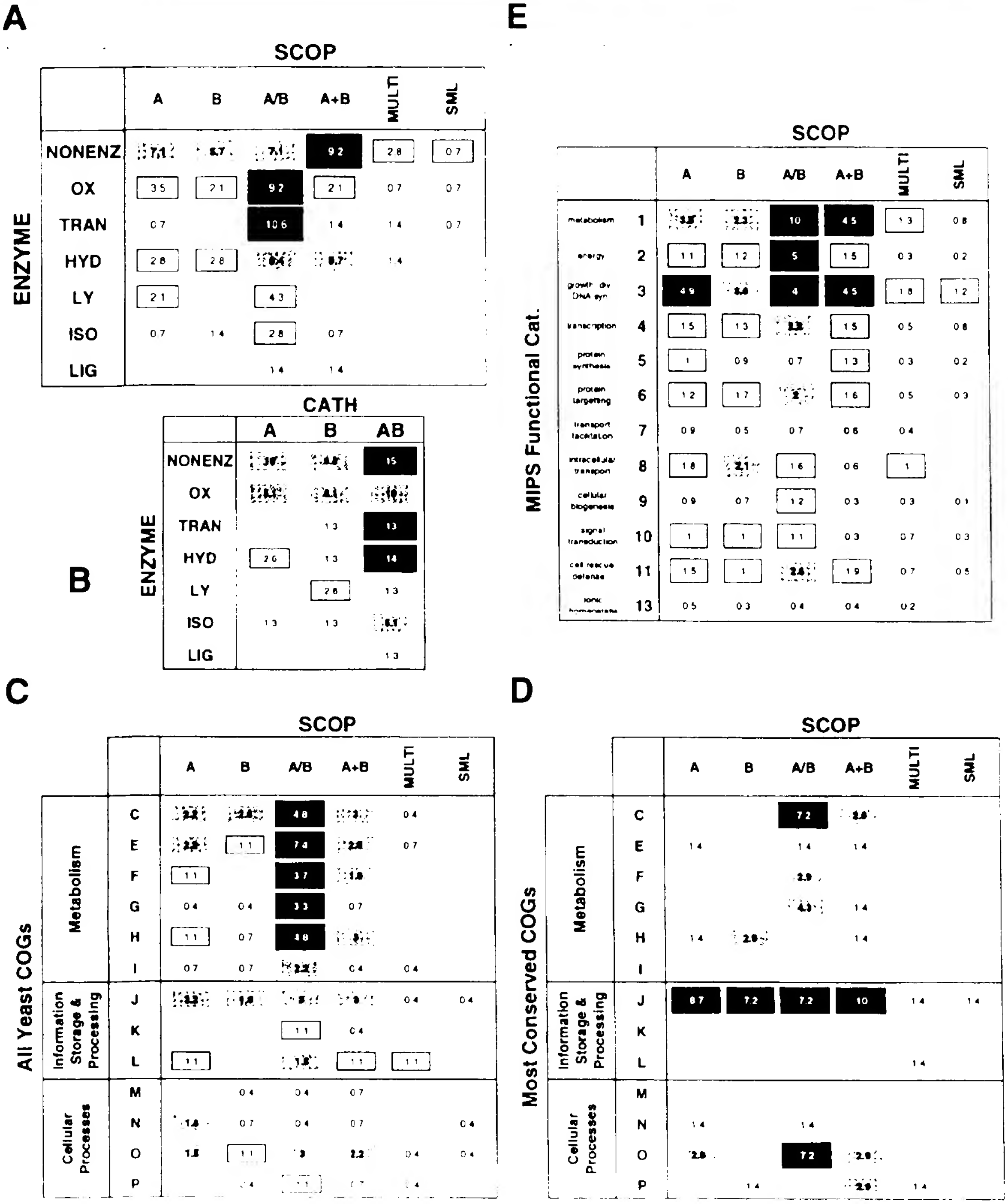


Figure 4. Structure-function classes in the yeast genome analyzed through a variety of classification schemes. This Figure shows the distribution of fold function combinations in the yeast genome as analyzed by a variety of different structure and functional classifications. Each of the Figures is a cross-tabulation of one structural classification scheme (on the column heads) *versus* a functional classification (row heads). (a) SCOP *versus* ENZYME; (b) CATH *versus* ENZYME; (c) SCOP *versus* COGs; (d) SCOP *versus* Most Conversed COGs; (e) SCOP *versus* MIPS Functional Catalogue. Each of the grid boxes gives the number of fold-function combinations within a structure-function class. This number is expressed as a percentage of the total number of combinations in the diagram to make the graphs readily comparable. The total number of combinations in each of the sub-figures is (a) 141, (b) 77, (c) 1207, (d) 120, and (e) 66. (a) and (e) are directly comparable with the cross tabulation in Table 2B for all of Swissprot. In (d) and (e), we employ the COGs scheme in exactly the same fashion as we did the ENZYME classification. We form combinations [illegible] SCOP [illegible] COG [illegible] with fold 2.26) and then we place these combi-

ners to create combinations with SCOP folds and then use the top number to create the functional classes [illegible] the diagram. For (e) we just use the 110 COGs that are present in all eight genomes in the current COGs analysis (E. coli, H. influenzae, H. pylori, M. genitalium, M. pneumoniae, Synechocystis, M. jannaschii and yeast).

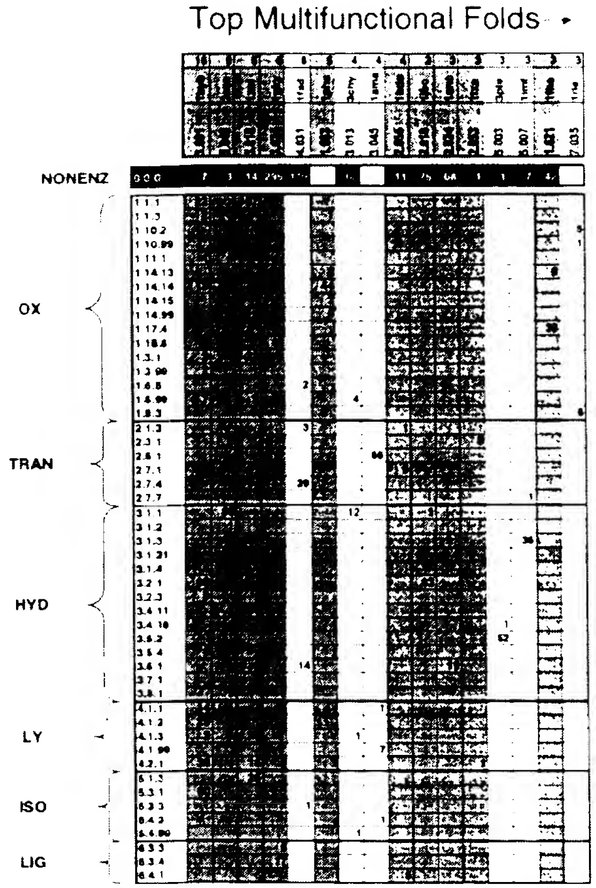


Figure 5. The most versatile folds. The functions associated with the 16 most versatile folds are shown. Values in the table denote the number of matches between a particular fold type in pdb95d (designated by its fold number in SCOP 1.35) and an enzyme category (represented by the first three components of the respective EC numbers). Here and in the following Tables the same parameters were used for matching as in Figure 2. The numbers in the top row indicate the number of functions a particular fold is associated with. The identifiers above the fold numbers are either PDB or SCOP identifiers of representative structures (the latter only if the PDB entry contains more than one domain or chain). (See the legend to Table 3 for the syntax of SCOP identifiers.) The first row in the table with the artificial 0.0.0 EC number shows the number of matches with non-enzymatic functions. Among the two all-alpha folds in the table, cytochrome P450 (1.063) is exclusively enzymatic, associated with five different enzyme functions, all related to cytochrome P450. Only one alpha + beta fold, ferredoxin (4.031), is present in the table, predominantly with matches with non-enzymatic ferredoxins, but also with enzymes in four different enzyme classes. In the multi-domain class, beta-lactamase/D-ala carboxypeptidase (5.003) has the most matches with penicillinase (EC number 3.5.2) and only one match with a non-enzyme, which also binds penicillin but has no enzymatic activity (Coque *et al.*, 1993). The class of small domains is represented only with one fold, membrane-bound rubredoxin-like (7.035), and has matches only with enzymes. It is possible that some proteins classified as "non-enzymes" may indeed be enzymes, missing the corresponding EC number. In this case, our analysis may be potentially useful in pointing to which non-enzymes may actually be enzymes.

Figure 4(e) shows another functional classification scheme, the MIPS Yeast functional catalogue (Mewes *et al.*, 1997). Unlike the COGs scheme, this has the advantage of being applicable to every yeast open reading frame (ORF). However, it has many more categories and about a third of the yeast ORFs are classified into multiple categories (sometimes five or more), making interpretation of the results a bit more ambiguous.

The most versatile folds and the most versatile functions

Returning to considerations of all of Swissprot, Figure 5 lists the 16 most versatile folds. The top five are the TIM-barrel, the alpha-beta hydrolase fold, the Rossmann fold, the P-loop containing NTP hydrolase fold, and the ferredoxin fold. Four of these are alpha/beta folds and one is alpha + beta. All five have non-enzymatic functions as well as five to 15 enzymatic ones. The most versatile folds include four all-beta and two all-alpha folds.

Figure 6 lists the 18 functions that have the most different folds associated with them, each having at least three associated folds. The most versatile functions are those of glycosidases and carboxylyases (3.2.1 and 4.2.1), which are associated with seven different fold types each, recruited from at least three different fold classes. The next two most versatile functions, the phosphoric monoester hydrolases and the linear monoester hydrolases (3.1.3 and 3.5.1), are associated with six different fold types each. Most of the versatile functions are associated with folds in completely different fold classes. This suggests that these enzymes developed independently, providing many examples of convergent evolution. In contrast, only three functions, all oxidoreductases, are associated with folds in a single class (last three rows in Figure 6). These folds are all alpha/beta, namely the TIM-barrel, Rossmann, and flavodoxin folds.

Specific functional convergences involving different folds

Even on the level of specificity of four-component EC numbers, several enzymatic functions are performed by unrelated structures. Figure 1 shows a dramatic example, two different carbonic anhydrases with the same EC number 4.2.1.1, but with clearly different structures (Kisker *et al.*, 1996). Table 3 shows further examples in a more systematic fashion. Most of these occur in different evolutionary lineages. For instance, the all-alpha vanadium chloroperoxidase occurs only in fungi, while the alpha/beta non-heme chloroperoxidase occurs only in prokaryotes. Another example is beta-glucanase. It has as many as three different structural representations, from three different fold classes. While it has an all-beta structure in *Bacillus subtilis*, it has an all-

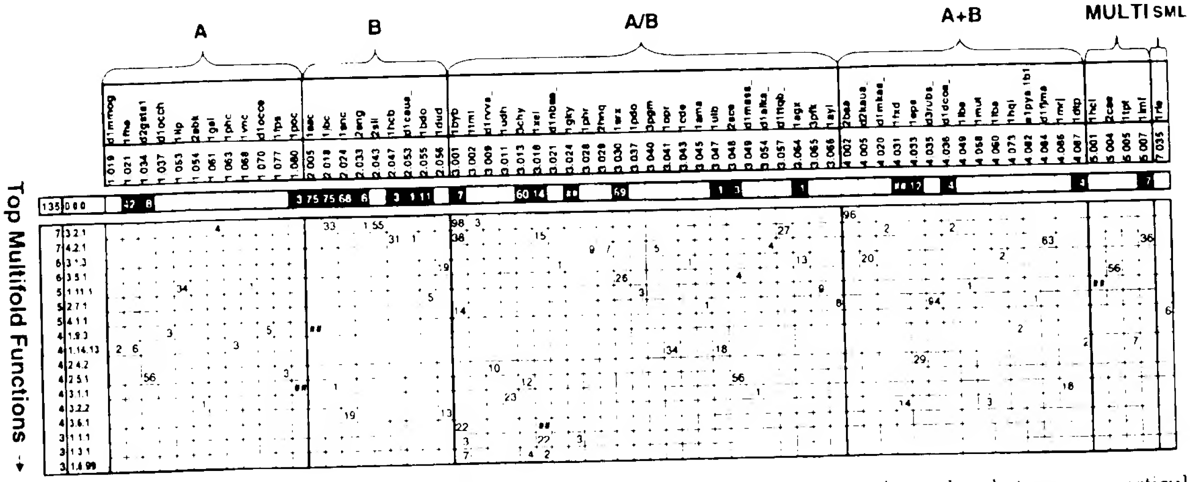


Figure 6. The most versatile functions. Values in the table denote the number of matches between a particular enzyme category (designated by the first three components of their EC numbers) and a SCOP 1.35 fold (designated by their fold numbers). This Figure follows the same conventions described in the legend to Figure 5. The rows are arranged in decreasing order according to the number of different folds with which they are associated (numbers shown in the first column). A hash (#) in any cell indicates that its value is greater than 99.

alpha variant in *Bacillus circulans*, and an alpha/beta structure in tobacco.

Specific functional divergences on same fold

Quite a number of SCOP domains each have sequence similarity with Swissprot proteins of different function. We separated these into cases in which the structural domain has similarity to proteins with different enzymatic functions only and those in which a domain shows homology to both enzymes and non-enzymes (Table 4A and B, respectively). Table 4A includes the well-known lactalbumin-lysozyme C similarity and the well-documented case of homology between an eye-lens structural protein and an enzyme (crystallin and gluthathione *S*-transferase; Cooper *et al.*, 1993; Qasba & Kumar, 1997). It includes several other notable divergences, such as the one between lysophospholipidase and galectin, and the one between an elastase and an antimicrobial protein (Morgan *et al.*, 1991). Remarkably, of the seven domains in this Table, three belong to the all-beta class.

"Multifunctionality" *versus e*-value

Figure 7 shows how the number of "multifunctional" domains, i.e. domains with sequence similarity to proteins with different functions, varies as the function of the stringency of the match score threshold. We used a minimal version of SCOP in which the structures in PDB were clustered into 990 representative domains (see the legend to Figure 7). The Figure shows how the percentage of domains that have sequence similarity to proteins

Table 3. Specific convergences

EC #	Enzymatic function	Fold #1	Dom #1	Swissprot 1	Fold #2	Dom #2	Swissprot 2
1.11.1.10	Chloroperoxidase	3.048.001	d1broa	PRXC_PSEPY	1.068.001	d1vnc	PRXC_CURIN
1.15.1.1	Superoxide dismutase	2.001.007	d1srda	SODI_ORYSA	4.023.001	d1mnga2	SODM_BACCA
3.1.3.48	Protein-tyrosine phosphatase	3.028.001	d1phr	PTPA_STRCO	3.029.001	d2hnp	PYP3_SCHPO
3.1.26.4	Ribonuclease h	3.038.003	d2rn2	RNH_ECOLI	3.039.001	d1ttr_	RNH_BPT4
3.2.1.4	Endoglucanase	1.061.001	d1cem_	GUN_BACSP	3.001.001	d1ecea	GUN_BACPO
3.2.1.8	Xylanase	2.018.001	d1yna_	XYN_TRIHA	3.001.001	d2exo_	XYNB_THENE
3.2.1.14	Endochitinase	3.001.001	d1hvq	CHIA_TOBAC	4.002.001	d2baa_	CHIX_PEA
3.2.1.73	Beta-glucanase*	3.001.001	d1ghr	GUB_NICPL	2.018.001	d1gbg	GUB_BACSU
3.2.1.73	Beta-glucanase	1.061.001	d1cem	GUB_BACCI			
3.2.1.91	Exoglucanase	2.018.001	d1cela	GUX1_TRIVI	3.002.001	d1cb2a	GUX3_AGABI
3.5.2.6	Beta-lactamase	5.003.001	d1btl	BLP4_PSEAE	4.083.001	d1bmc	BLAB_BACCE
4.2.1.1	Carbonic anhydrase	2.053.001	d1thja	CAH_METTE	2.047.001	d2cba	CAHZ_BRARE
[illegible]	[illegible]	[illegible]	[illegible]	MIP_TRYCR	2.041.001	d2cpl	CYPR_DROME

[illegible] script syntax: d1pdbCN, where 1pdb is a PDB code, C is a chain identifier and N describes the position of the first, second [illegible] domain in the chain. Thus d1gata1 is the first domain in the A chain of 1GAT.

Table 4. Specific divergences

A. *Two different enzymatic functions*

SCOP domain	Fold number	Swissprot 1	FC num 1	Function 1	Swissprot 2	EC num 2	Function 2
d2abk	1.001.054.001.001.001	END3_ECOLI	4.2.99.18	Endonuclease III	GTMR_METTF	3.2.2.-	Possible G-T mismatches repair enzyme
d1bdo	1.002.055.001.001.001	BCCP_ECOLI	6.4.1.2	Biotin carboxyl carrier protein of acetyl-Coa carboxylase	BCCP_PROFR	2.1.3.1	Biotin carboxyl carrier protein of methylmalonyl-CoA carboxyl-transferase
d1dhpa	1.003.001.003.001.004	NPL_ECOLI	4.1.3.3	*N*-Acetylneuraminate lyase subunit	DAPA_BACSU	4.2.1.52	Dihydrodipicolinate synthase
d1hdca	1.003.018.001.002.005	ENTA_FCOLI	1.3.1.28	2,3 Dihydro-2,3 dihydroxy-benzoate dehydrogenase	ADHI_DROMO	1.1.1.1	Alcohol dehydrogenase 1
d1nipa	1.003.024.001.005.003	BCHL_RHOCA	1.3.1.33	Protochlorophillide reductase 33 kD subunit	NIFH_THIFE	1.18.6.1	Nitrogenase iron protein
d1gara	1.003.043.001.001.001	PUR3_YEAST	2.1.2.2	Phosphoribosylglycinamide formyltransferase	PURU_CORSP	3.5.1.10	Formyltetrahydrofolate deformylase
d2dkb_	1.003.045.001.003.001	OAT_RAT	2.6.1.13	Ornithine aminotransferase precursor	GSAB_BACSU	5.4.3.8	Glutamate-1-semialdehyde 2,1-aminomutase 2
d1ede	1.003.048.001.003.001	DMPD_PSEPU	3.1.1.-	2-Hydroxymuconic semialdehyde hydrolase	HALO_XANAU	3.8.1.5	Haloalkane dehalogenase
d1fua	1.003.053.001.001.001	ARAD_ECOLI	5.1.3.4	L-Ribulose-5-phosphate 4-epimerase	FUCA_ECOLI	4.1.2.17	L-Fuculose phosphate aldolase
d1lmn	1.004.002.001.002.010	LCA_RAT	2.4.1.22	Alpha-lactalbumin precursor	LYC1_PIG	3.2.1.17	Lysozyme C-1
d1frva	1.005.015.001.001.001	FRHG_METVO	1.12.99.1	Coenzyme F420 hydrogenase gamma subunit	MBHS_AZOCH	1.18.99.1	Uptake hydrogenase small subunit precursor

B. *Enzyme and non-enzyme*

SCOP domain	Fold number	Swissprot 1	Enzymatic function	EC number	Swissprot 2	Non-enzymatic function
d1gsq_1	1.001.034.001.001.007	GTS2_MANSE	Glutathione *S*-transferase 2	2.5.1.18	SC11_OMMSL	*S*-Crystallin SL11 (major lens polypeptide)
d1lcl	1.002.018.001.003.003	LPPL_HUMAN	Eosinophil lysophospholipase	3.1.1.5	LEG7_RAT	Galectin-7
d1brbe	1.002.029.001.002.003	CFAD_RAT	Endogenous vascular elastase	3.4.21.46	CAP7_HUMAN	Azurocidin (antimicrobial, heparin-binding protein)
d1mup_	1.002.039.001.001.007	PGHD_HUMAN	Prostaglandin-D synthase	5.3.99.2	LACC_CANFA	Beta-lactoglobulin III
_d1mup	1.002.039.001.001.007				QSP_CHICK	Quiescence-specific protein
d2hhma_	1.005.007.001.002.001	MYOP_XENLA	Inositol mono-phosphatase	3.1.3.25	SUHB_ECOLI	Extragenic suppressor protein SUHB
d2hhma	1.005.007.001.002.001	STRO_STRGR	DTDP-glucose synthase	2.7.7.24		
d1isua	1.007.029.001.001.001	IRO_THIFE	Iron oxidase precursor (FE(II) oxidase)	1.16.3.-	HPIT_RHOTE	High potential iron-sulfur protein (HIPIP)

List of SCOP domains that are each homologous to several Swissprot proteins with significantly different function. In A, the domains homologous to proteins with different (in the last three component of EC numbers) enzymatic functions are listed. In most cases, the enzymatic functions remain analogous, as reflected in the names of the enzymes. B lists the domains homologous to proteins with both enzymatic and non-enzymatic functions. (See Table 3 for the SCOP domain syntax.)

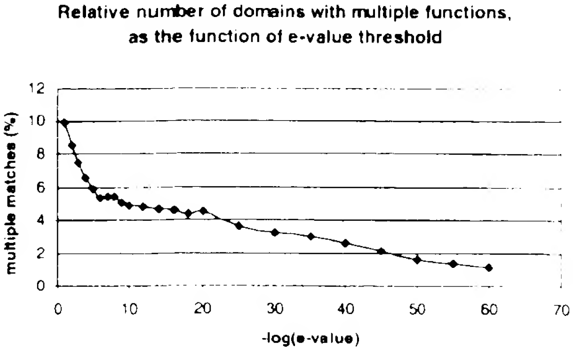


Figure 7. Multi-functionality *versus* *e*-value threshold. The graph shows how the percentage number of multi-functional enzymatic domains varies as the function of the *e*-value threshold. A multi-functional domain occurs when a particular domain in SCOP matches domains in Swissprot with different enzymatic function. For these calculations, we had to use a more minimal version of SCOP than the pdb95d dataset referred to in the methods to prevent double matches, i.e. two SCOP domains matching a single Swissprot domain. The construction of this minimal SCOP was described previously (Gerstein, 1998a). Basically, all the domains in SCOP were clustered *via* a multi-linkage approach into 990 representative domains, such that no two domains matched each other with a FastA *e*-value better than 0.01.

with different functions (in terms of three-component EC numbers) varies with sequence similarity. This decreases approximately monotonically as a function of the exponent of the *e*-value threshold. Interestingly, there is a breaking point around log (*e*-value) = −5, as the sharply decreasing number of functions slows down and the matches reach the level of biological significance.

Our graph can be loosely compared with the classic graph by Chothia & Lesk (1986) showing the relation of similarity in structure to that in sequence. It roughly shows the chance of functional similarity (or more precisely the chance of functional difference) with a given level of sequence similarity between an enzyme and a protein of unknown function. For example, with an *e*-value of 10^{-10}, there is only an 5% chance that an unknown protein homologous to a certain enzyme has in fact a different function. Moreover, our graph is in excellent agreement with the findings by Russell *et al.* (1998) who also found that the proportion of homologues with different functions is around 10%. This shows that there is a low chance that a single-domain protein, highly homologous to a known enzyme, has a different

Overview

We have investigated the relationship between the structure and function of proteins by comparing functionally characterized enzymes in Swissprot with structurally characterized domains in SCOP. It is a timely subject, as the number of three-dimensional protein structures is increasing rapidly and the recent completion of several microbial genomes highlights the need for functional characterization of the gene products and identification of enzymes participating in metabolic pathways (Koonin *et al.*, 1998).

We tried to be as objective and as unbiased as possible, taking only enzymes with a single assigned function and only single-domain matches. We ignored Swissprot proteins with dubious or unknown function, or with incomplete sequence. Given these criteria, several tendencies are clear. The alpha/beta folds tend to be enzymes. The all-alpha folds tend to be non-enzymes and the all-beta and alpha + beta folds tend to have a more even distribution between enzymes and non-enzymes.

Our analysis of proteins from yeast and *E. coli* has shown that the functional distribution of folds does not differ greatly from the whole of Swissprot. *E. coli*, however, appears to have somewhat more alpha/beta enzymes and less non-enzymes.

Functional assignment complexities

We identified four specific complexities in our functional assignment worth mentioning.

Firstly, there is not always a one-to-one relationship between gene protein and reaction (Riley, 1998). An enzyme can have two functions, or two polypeptides from two different genes can oligomerize to perform a single function. It might be that some of the fold-functions combinations in Figure 2 occur together in multi-domain proteins (which otherwise were not the subject of this survey). An exhaustive screening revealed that only four pairs of folds in Figure 2 were present concurrently in multi-domain proteins. Each of these reduced by one the number of independent fold-function combinations. (The four pairs were as follows, with one representative Swissprot protein in each category, EC numbers in parentheses, and then SCOP fold numbers: PTAA_ECOLI (2.7.1) has 4.049 and 2.055 folds, TRP_COPCI (4.2.1) has 3.057 and 4.005 folds, URE1_HELFE (3.5.1) has 4.005 and 2.056 folds, while XYNA_RUMFL (3.2.1) has 2.018 and 3.001 folds.)

Secondly, the functions associated with similar structures often turn out to be analogous, even if they show significant difference in their EC numbers. For example, acetyl-CoA carboxylase and

[illegible] EC classification numbers [illegible] and [illegible] respectively)

Thirdly, there are clearly some drawbacks to the EC system. The EC system is a classification of reactions, not underlying biochemical mechanisms. An enzyme classification system based explicitly on reaction mechanism (e.g. "involves pyridoxal phosphate" or "involves Ser as a nucleophile") might also prove interesting to compare with protein structure. Alternatively, one based on pathways might be worthwhile since, as pointed out by Martin *et al.* (1998), "it may be that more significant relationships occur within pathways, where the substrate is successively transferred from enzyme to enzyme along the pathway, requiring similar binding sites at each stage".

Finally, in all of Swissprot the majority of the 101 folds with only non-enzymatic functions probably have several functions, but we were not able to consider them separately here, lacking a general protein function classification system for non-enzymes. Such a system is not easy to derive. For instance, if we took only the first three words of all the description lines in Swissprot, we would end up with about 10,000 different protein functions (besides enzymes). An approximate solution to this problem is offered by a recent work that has classified 81 % of Swissprot into one of three broad categories in an automated fashion (Tamames *et al.*, 1997). However, one way we did tackle this problem was by focussing on the yeast genome for which there are a number of overall functional classification systems. This work showed that the preferred association of folds with certain functions occurs for non-enzymes as well as enzymes. Furthermore, the results for the highly conserved COGs would be expected to be exactly the same in other genomes.

Biases

Our results are undoubtedly affected to some degree by the biases inherent in the databanks, e.g. towards mammalian, medically relevant proteins and towards proteins that easily crystallize. Such biases probably result in the higher representation of enzymes in the structural databases, in the PDB and therefore in SCOP. This might be the cause of the higher occurrence of alpha/beta proteins in our tables and the higher density of matches in this class.

One interesting question related to biases is whether looking only at individual genomes instead of the whole database will give different results. Our results for yeast suggest that it is not necessarily the case.

Comparison with Martin *et al.* (1998)

Martin *et al.* (1998) performed a similar analysis to the one described here. One of the conclusions of their careful study was that there was no relationship between the top-level CATH classification and the top-level EC class. This seems to be at odds with our results. However, we have found the conclusions to be consistent. There are a number of reasons for this.

Firstly, Martin *et al.* (1998) tabulate statistics on only the proteins in the PDB. They found a clear alpha/beta preference for proteins in the oxidoreductase, transferase, and hydrolase categories (EC 1-3), but for the lyase, isomerase, and ligase categories (EC 4-6) they observe different tendencies. However, they did not have sufficient counts to establish statistical significance for this latter finding. (This is basically what we observe in Figure 4(b).) Because in our analysis we use all of Swissprot and we tabulate our statistics a little differently (in terms of combinations), we get more "counts" than Martin *et al.* (1998). Thus, we are able to argue that the different distribution of fold-function combinations observed for lyases, isomerases, and ligases are significant. This is borne out by the chi-squared statistics at the end of Table 2.

Secondly, Martin *et al.* "no-relationship" conclusion applies only to comparisons between the different enzyme classes. However, we find our largest differences when comparing non-enzymes to enzymes and also comparing between the various types of non-enzymes.

Finally, the CATH classification that Martin *et al.* use has only three classes in its top-most level. In contrast, SCOP has six top classes (Table 1). While this larger number of categories does tend to degrade our statistics somewhat, it also highlights some differences that cannot be observed in terms of the CATH classes alone, e.g. we find clear differences between alpha + beta and alpha/beta proteins and also between small proteins and all others.

Apparently high occurrence of convergent evolution

Note that the table in Figure 2 is not square: it has more folds than functions. This shape leads to a number of interesting conclusions. The 331 fold-function combinations we observe for 229 folds and 92 functions imply that there are 1.2 functions per fold and 3.6 folds per function. However, these numbers are somewhat skewed by the large number of folds (101) associated only with the single non-enzymatic function. If we exclude these, we get 128 "enzyme-related" folds, which are, in turn, associated with 230 (= 331 − 101) different fold-function combinations. This implies that for the enzyme-related folds there are on average 1.8 functions per fold and 2.5 folds per function (230/128 and 230/92). The larger number of folds per function than functions per fold seems to suggest that nature tends to reinvent an enzymatic function (i.e. convergent evolution) more often than modify an already existing one (i.e. functional divergence).

How can we explain this? Firstly, 1.8 is a lower estimation for the number of functions per fold as the non-enzymatic functions were bundled into one group here. Secondly, there are several examples of functional divergence for a fold within one three-component enzyme category that are not reflected in our Tables. For instance, the 1.1.1 category has 248 different enzymes, which all share the same fold. Thirdly, the results in this paper were derived from databases comprised of data from several organisms. It is quite possible that within one organism, functional divergence is more prevalent than convergent evolution.

Superfolds and superfunctions

Are functions more diverse for the more common folds? To some degree this brings up a "chicken-and-egg" issue. Do folds have more functions because they occur more often or is it the other way around? The commonness of a fold is often quantified by the number of non-homologous sequence families accommodated by the fold, and folds accommodating many families of diverse sequences have been dubbed "superfolds" (Orengo *et al.*, 1993). We find that there seems to be a loose connection between the number of diverse sequence families associated with a particular fold (in SCOP) and the functional diversity of that fold. For instance, the top superfold is the TIM-barrel; it also has the most functions associated with it (15 different enzymatic functions as shown in Figure 4). On the other hand, there are exceptions: the alpha/beta hydrolases and the Rossmann fold are both associated with 22 sequence families in SCOP, but while the former has eight different enzymatic functions, the latter has only three.

Finally, while there is a high incidence of particular functions with many folds ("superfunctions"), as well as folds with many functions, the distribution of superfunctions appears to be more uniform and less concentrated on a few exceptionally versatile individuals than is the case for folds. That is, comparing Figures 3 and 4 one can see that the top nine most versatile functions are associated with five to seven folds while the top nine most versatile folds carry out from six to as many as 16 functions. This last value is for the TIM-barrel and underscores the uniqueness of this fold as a generic scaffold (see Figure 1 for an illustration of this fold).

Why folds are associated with functions: chemistry *versus* history

Why is a certain fold chosen to carry out a particular function? It is, of course not possible to answer this question definitively at present. However, there are two broad themes that emerge from our analysis. The first is favorable chemistry. Per

more chemically suitable. This could be the situation for the ribosomal proteins (and is borne out by the results of Figure 4(d)).

Materials and Methods

Sequence matching to swissprot

All the protein sequences in Swissprot 35 were compared with all the protein domain sequences in SCOP 1.35 by standard database search programs (WU-BLAST; Altschul *et al.*, 1990). The following five criteria were used in the searches: (1) At least three of the four components of the EC number are assigned in the DE line of the Swissprot entries. (2) Fragments in Swissprot were excluded (this affected about 10% of the entries). (3) For WU-BLAST searches an *e*-value threshold of 0.0001 was used, unless stated otherwise. (4) Only "monoenzymes", i.e. proteins with only one enzymatic function, were considered. This excluded less than 0.5% of the Swissprot enzymes. (5) Only single-domain matches with Swissprot proteins were taken into consideration. This means those proteins that had a match with a SCOP domain covering most of the Swissprot protein. Specifically, we required that less than 100 amino acid residues be left uncovered in the Swissprot entry by a match. We are aware that this is only an approximation, as there are domains with less than 100 amino acid residues; however, it is considerably less than the average length of a SCOP domain (163 residues) and seems to be a reasonable threshold in an automated approach.

All the searches were repeated using FASTA with an *e*-value threshold of 0.01 (Pearson, 1998; Pearson & Lipman, 1988). The results obtained by the two different comparison programs were in agreement with each other. That is, the FASTA searches did not result in any new combinations of folds and enzymatic functions (a new dot in Figure 1), and therefore are not shown.

Sequence matching to the yeast genome

To get as great a coverage of the yeast genome as possible, we did a sequence comparison for *just* Figure 4 using an altered protocol. We first ran the PDB against the yeast genome using FASTA and kept all matches with a better than 0.01 *e*-value (Pearson, 1998; Pearson & Lipman, 1988). Then, to increase our number of matches further we used the PSI-blast program (Altschul *et al.*, 1997). This program is somewhat more complex to run than FASTA, involving embedding the yeast genome in NRDB and running PDB query sequences against it in an iterative fashion, adding the matches found at each round to a growing profile. We used the PSI-blast parameters adapted from Teichmann *et al.* (1998): an *e*-value threshold of 0.0005 to include matches in the profile and iteration of up to 30 times or to convergence. We did not continuously parse the output and accepted matches at the final iteration that had *E*-value scores better than 0.0001. The number of iteration to [illegible]

[illegible] and its function reflects a particular accident that took place at the beginning of cellular evolution. However, once this choice was made it was impossible to undo even if other folds would be

finds many more matches than either FASTA or WU-BLAST. However it has problems with certain small and compositionally biased proteins. We used FASTA for these and also tried to remove compositional bias through running the SEG program with standard parameters (Wootton & Federhen, 1996).

How the structural classifications were used: SCOP and CATH

SCOP hierarchically clusters all the domains in the PDB database, assigning a five-component number to each domain (Murzin *et al.*, 1995). The first component in the SCOP numbers denotes the structural class to which the domain in question belongs. The second component of the SCOP numbers designates the fold type of the domain. There are altogether 361 different fold types in SCOP 1.35. The six SCOP classes used in this survey are listed in Table 1B.

In this study, a 95% non-redundant subset of SCOP was used, i.e. all pairs of domains had less than 95% sequence homology. This set is denoted pdb95d and is available from the SCOP website (scop.mrc-lmb.cam.ac.uk). We used version 1.35, which had 2314 protein domains. (The yeast analysis used a more recent version of SCOP, 1.38, which had 3206 domains.)

The CATH classification classifies structures in analogous fashion to SCOP (Orengo *et al.*, 1997). However, the exact structure of the classification is not the same, with an additional architecture level inserted between the top-level class and the fold-level. In our use of the classification, we created a limited mapping table that associated each SCOP domain in pdb95d with its corresponding classification in CATH 1.4. This was not always possible to do unambiguously. As a result, we left out the ambiguous matches from the statistics.

How the functional classifications were used: ENZYME, COGS, and MIPS

The EC numbers of enzymes are composed of four components (Barrett, 1997). (1) The first component shows to which of the six main divisions the enzyme belongs. (2) The second figure indicates the subclass (referring to the donor in oxidoreductases or the group transferred in transferases, or the affected bond in hydrolases, lyases or ligases). (3) The third figure indicates the sub-subclass (e.g. indicating the type of acceptor in oxidoreductases). (4) The fourth figure gives the serial number of the enzyme in its sub-subclass. The six main divisions are listed in Table 1A.

In the analysis of all of Swissprot, when we counted the number of non-enzymatic matches, all the proteins called 'HYPOTHETICAL' and all the proteins having an '-ase' word ending but lacking an EC number in their description were excluded, because of their functional ambiguity. For relating the sequence matches of the yeast genome to the EC system, we used essentially the same criteria as we did for all of Swissprot (see above): single-domain, monoenzyme matches with at least a three-component EC number.

The COGs and especially the MIPS classifications are a bit more complex than the EC system in that they include non-enzymes as well as enzymes (Tatusov *et al.*, 1997; Koonin *et al.*, 1998; Mewes *et al.*, 1997). They often associate multiple functions or roles to a given yeast ORF. This happens for more than a third of the yeast ORFs with MIPS. In this case, if we could clearly show a PDB match was associated with a single functional domain we made only that pairing. Otherwise we associated all the functions assigned to a given PDB match to its respective fold.

Availability of results over the internet

A number of detailed tables relevant to our study will be made available over the Internet at http://bioinfo.mbb.yale.edu/genome/foldfunc, in particular, a "clickable" version of Figure 1 and large data files giving all the fold assignment and fold-function combinations for Swissprot and yeast.

Acknowledgments

We thank the Donaghue Foundation and the ONR for financial support (grant N000149710725). We thank Ted Johnson for help with the minimal version of the SCOP database.

References

Altschul, S., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P. & Selley, J. N. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucl. Acids Res.* **26**, 304-308.

Bairoch, A. (1996). The ENZYME data bank in 1995. *Nucl. Acids Res.* **24**, 221-222.

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* **26**, 38-42.

Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucl. Acids Res.* **25**, 217-221.

Barrett, A. J. (1997). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *Eur. J. Biochem.* **250**, 1-6.

Bork, P. & Eisenberg, D. (1998). Deriving biological knowledge from genomic sequences. *Curr. Opin. Struct. Biol.* **8**, 331-332.

Bork, P. & Koonin, E. V. (1998). Predicting functions from protein sequences-where are the bottlenecks? *Nature Genet.* **18**, 313-318.

Bork, P., Sander, C. & Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Sci.* **2**, 31-40.

Bork, P., Ouzounis, C. & Sander, C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4**, 393-403.

Chen, L., DeVries, A. L. & Cheng, C. H. (1997). Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc. Natl Acad. Sci. USA*, **94**, 3817-3822.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.

Cooper, D. L., Isola, N. R., Stevenson, K. & Baptist, E. W. (1993). Members of the ALDH gene family are lens and corneal crystallins. *Advan. Exp. Med. Biol.* **328**, 169-179.

Coque, J. J., Liras, P. & Martin, J. F. (1993). Genes for a beta-lactamase, a penicillin-binding protein and a transmembrane protein are clustered with the cephamycin biosynthetic genes in *Nocardia lactamdurans*. *EMBO J.* **12**, 631-639.

Corpet, F., Gouzy, J. & Kahn, D. (1998). The ProDom database of protein domain families. *Nucl. Acids Res.* **26**, 323-326.

des, Jardins M., Karp, P. D., Krummenacker, M., Lee, T. J. & Ouzounis, C. A. (1997). Prediction of enzyme classification from protein sequence without the use of sequence similarity. *ISMB*, **5**, 92-99.

Doolittle, R. F. (1994). Convergent evolution: the need to be explicit. *Trends Biochem. Sci.* **19**, 15-18.

Fabian, P., Murvai, J., Hatsagi, Z., Vlahovicek, K., Hegyi, H. & Pongor, S. (1997). The SBASE protein domain library, release 5.0: a collection of annotated protein sequence segments. *Nucl. Acids Res.* **25**, 240-243.

Frishman, D. & Mewes, H.-W. (1997). Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4**, 626-628.

Galperin, M. Y., Walker, D. R. & Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* **8**, 779-790.

Gerstein, M. (1997). A structural census of genomes: comparing eukaryotic, bacterial and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562-576.

Gerstein, M. (1998a). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Design*, **3**, 497-512.

Gerstein, M. (1998b). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* **33**, 518-534.

Gerstein, M. & Hegyi, H. (1998). Comparing microbial genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* **22**, 277-304.

Gerstein, M. & Levitt, M. (1997). A structural census of the current population of protein sequences. *Proc. Natl Acad. Sci. USA*, **94**, 11911-11916.

Hellinga, H. W. (1997). Rational protein design: combining theory and experiment. *Proc. Natl Acad. Sci. USA*, **94**, 10015-10017.

Hellinga, H. W. (1998). Computational protein engineering. *Nature Struct. Biol.* **5**, 525-527.

Henikoff, S., Pietrokovski, S. & Henikoff, J. G. (1998). Superior performance in protein homology detection with the Blocks Database servers. *Nucl. Acids Res.* **26**, 309-312.

[illegible] & Sander, C. [illegible] space with Dali FSSP. *Nucl. Acids Res.* **26**, 316-319.

Ibba, M., Bono, J. L., Rosa, P. A. & Soll, D. (1997a). Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. *Proc. Natl Acad. Sci. USA*, **94**, 14383-14388.

Ibba, M., Morgan, S., Curnow, A. W., Pridmore, D. R., Vothknecht, U. C., Gardner, W., Lin, W., Woese, C. R. & Soll, D. (1997b). A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science*, **278**, 1119-1122.

Karp, P. (1998). What we do not know about sequence analysis and sequence databases. *Bioinformatics*, **14**, 753-754.

Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. (1998). EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* **26**, 50-53.

Kisker, C., Schindelin, H., Alber, B. E., Ferry, J. G. & Rees, D. C. (1996). A left-hand beta-helix revealed by the crystal structure of a carbonic anhydrase from the archaeon *Methanosarcina thermophila*. *EMBO J.* **15**, 2323-2330.

Koonin, E. V. & Galperin, M. Y. (1997). Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr. Opin. Genet. Dev.* **7**, 757-763.

Koonin, E. V. & Tatusov, R. L. (1994). Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J. Mol. Biol.* **244**, 125-132.

Koonin, E. V., Tatusov, R. L. & Galperin, M. Y. (1998). Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* **8**, 355-363.

Kraulis, P. J. (1991). MOLSCRIPT-a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C. & Thornton, J. M. (1998). Protein folds and functions. *Structure*, **6**, 875-884.

Marvin, J. S., Corcoran, E. E., Hattangadi, N. A., Zhang, J. V., Gere, S. A. & Hellinga, H. W. (1997). The rational design of allosteric interactions in a monomeric protein and its applications to the construction of biosensors. *Proc. Natl Acad. Sci. USA*, **94**, 4366-4371.

Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F. & Zollner, A. (1997). Overview of the yeast genome. *Nature*, **387**, 7-65.

Morgan, J. G., Sukiennicki, T., Pereira, H. A., Spitznagel, J. K., Guerra, M. E. & Larrick, J. W. (1991). Cloning of the cDNA for the serine protease homolog CAP37/azurocidin, a microbicidal and chemotactic protein from human granulocytes. *J. Immunol.* **147**, 3210-3214.

Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & [illegible]

[illegible] Michie, A. D., Jones, S., [illegible] Swindells, M. B. & Thornton, J. M. (1997). CATH a hierarchic classification of protein domain structures. *Structure*, **5**, 1093-1108.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227-259

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71-84.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA*, **85**, 2444-2448.

Qasba, P. K. & Kumar, S. (1997). Molecular divergence of lysozymes and alpha-lactalbumin. *Crit. Rev. Biochem. Mol. Biol.* **32**, 255-306.

Riley, M. (1997). Genes and proteins of *Escherichia coli* K-12 (GenProtEC). *Nucl. Acids Res.* **25**, 51-52.

Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211 1227.

Russell, R. B., Sasieni, P. D. & Sternberg, M. J. E. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903-918.

Seery, L. T., Nestor, P. V. & FitzGerald, G. A. (1998). Molecular evolution of the aldo-keto reductase gene superfamily. *J. Mol. Evol.* **46**, 139-146.

Selkov, E., Galimova, M., Goryanin, I., Gretchkin, Y., Ivanova, N., Komarov, Y., Maltsev, N., Mikhailova, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L. & Selkov, E., Jr (1997). The metabolic pathway collection: an update. *Nucl. Acids Res.* **25**, 37-38.

Sonnhammer, E., Eddy, S. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* **28**, 405-420.

Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66-73.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, **278**, 631-637.

Teichmann, S., Park, J. & Chothia, C. (1998). Structural assignments to the proteins of *Mycoplasma genitalium* show that they have been formed by extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA*, **95**, 14658-14663.

Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554-571.

Edited by G. von Heijne

(Received 16 November 1998; received in revised form 1 March 1999; accepted 1 March 1999)

Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins

Hedi Hegyi and Mark Gerstein[1]

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA

Annotation transfer is a principal process in genome annotation. It involves "transferring" structural and functional annotation to uncharacterized open reading frames (ORFs) in a newly completed genome from experimentally characterized proteins similar in sequence. To prevent errors in genome annotation, it is important that this process be robust and statistically well-characterized, especially with regard to how it depends on the degree of sequence similarity. Previously, we and others have analyzed annotation transfer in single-domain proteins. Multi-domain proteins, which make up the bulk of the ORFs in eukaryotic genomes, present more complex issues in functional conservation. Here we present a large-scale survey of annotation transfer in these proteins, using scop superfamilies to define domain folds and a thesaurus based on SWISS-PROT keywords to define functional categories. Our survey reveals that multi-domain proteins have significantly less functional conservation than single-domain ones, except when they share the exact same combination of domain folds. In particular, we find that for multi-domain proteins, approximate function can be accurately transferred with only 35% certainty for pairs of proteins sharing one structural superfamily. In contrast, this value is 67% for pairs of single-domain proteins sharing the same structural superfamily. On the other hand, if two multi-domain proteins contain the same combination of two structural superfamilies the probability of their sharing the same function increases to 80% in the case of complete coverage along the full length of both proteins, this value increases further to > 90%. Moreover, we found that only 70 of the current total of 455 structural superfamilies are found in both single and multi-domain proteins and only 14 of these were associated with the same function in both categories of proteins. We also investigated the degree to which function could be transferred between pairs of multi-domain proteins with respect to the degree of sequence similarity between them, finding that functional divergence at a given amount of sequence similarity is always about two-fold greater for pairs of multi-domain proteins (sharing similarity over a single domain) in comparison to pairs of single-domain ones, though the overall shape of the relationship is quite similar. Further information is available at http://partslist.org/func or http://bioinfo.mbb.yale.edu/partslist/func.

The ultimate goal of the genome projects is to determine the structure and function of all the newly identified gene products. Fundamentally, this will be carried out via annotation transfer, transferring the structural and functional annotation from an experimentally characterized protein (as in a model organism such as *Escherichia coli*) to a predicted protein in a newly sequenced genome that shares similarity in sequence. The degree of annotation transferred will depend on the degree of sequence similarity. This process is shown schematically in Figure 1. In this paper, we aim to address this major question in bioinformatics, specifically focusing on multi-domain proteins, as they make up the bulk of the proteome in eukaryotic organisms (Gerstein 1998).

Our work is a direct outgrowth of two previous analyses of ours that concentrated on single-domain proteins. In an earlier paper, we found that the different structural classes of the scop classification system have different propensities to carry out certain types of function (Hegyi and Gerstein 1999). In particular, while the alpha/beta folds were disproportionately associated with enzymes and all-alpha and small folds with non-enzymes, the alpha + beta structures had an equal tendency for both enzymatic and non-enzymatic functions. Wilson et al. (2000) compared a large number of protein domains to one another in a pair-wise fashion with respect to similarities in sequence, structure, and function. Using a hybrid functional classification scheme merging the ENZYME and FlyBase systems (Gelbart et al. 1997; Bairoch 2000), they found that precise function is not conserved below 30–40% identity, although the broad functional class is usually preserved for sequence identities as low as 20–25%, given that the sequences have the same fold. Their survey also reinforced the previously established general exponential relationship between structural and sequence similarity (Chothia and Lesk 1986).

Other Work on Establishing Relationships between Sequence, Structure, and Function

Several other groups have studied the relationship between sequence, structure, and function in detail, attempting to determine the extent to which functional transference between matching proteins is feasible (Shah and Hunger 1997; Martin et al. 1998; Thornton et al. 1999, 2000; Zhang et al. 1999; Shapiro and Harris 2000; Todd et al. 2001). Orengo et al.

[1]Corresponding author.

2001 Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

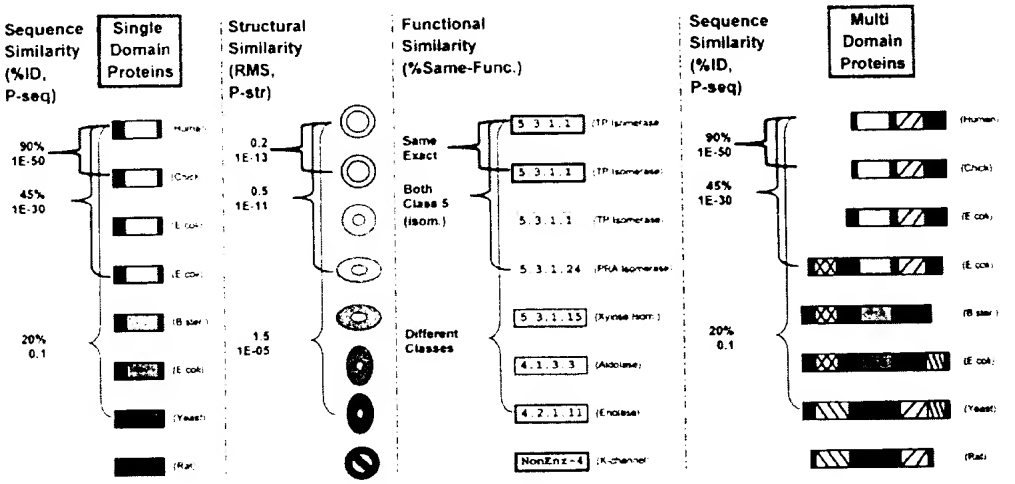


Figure 1 Schematic illustrating annotation transfer. This figure illustrates the process of annotation transfer for a group of hypothetical TIM barrel proteins. The *leftmost* panel represents sequence comparisons between idealized barrel domains from a number of organisms. The next panel shows analogous results for structural comparison, and the panel after that, functional comparison. The *rightmost* panel represents sequence comparisons between idealized multi-domain proteins that match over a single domain, the subject of much of this paper.

Pawlowski et al. (2000) studied the relationship between sequence and functional similarity in the twilight zone of 10%–15% sequence similarity and found a clear correlation between the two, with functional similarity based on the E.C. classification of enzymes.

Russell et al. (1997) analyzed binding sites in proteins with similar 3D structures and estimated that 90% of new remote homolog have common binding sites and similar functions. Eisenstein et al. (2000) evaluated the first results from the structural genomics projects and found that in many instances the protein structure itself offers an important clue to its biological function. Stawiski et al. (2000) found that function could be predicted rather successfully for just the proteases. Devos and Valencia (2000) presented a critical view of function transference between similar sequences, highlighting the limitations of this process due to errors in databases and the inherent complexity of the relationship between protein sequence-structure and function that does not allow "simplistic interpretations." They also found that binding sites are the least conserved features between related proteins while the catalytic activity of enzymes is the most conserved one.

Multi-Domain Proteins with Divergent Functions: How Common?

Most of these previous investigations focused on single-domain proteins or did not distinguish between single- and multi-domain ones. It is not clear how the multi-domain proteins with various functions behave with respect to functional conservation; namely, whether they are more or less conserved than their single-domain counterparts. In particular, as shown in Figure 1, if one multi-domain protein shares a single domain fold with another one, it is not clear the degree to which the functional conservation of these proteins is conserved of the SH3-domain (scop superfamily identifier 2.24.2) and the P-loop containing NTP hydrolase (3.29.1). While in higher organisms this combination is associated with presynaptic and tumor suppressor functions (SWISS-PROT names SP02_HUMAN and DLGI_DROME, respectively), in the lower Dictyostelium it was found in myosin (MYSP_DICDI). Another example is the combination of the FAD/NAD(P)-binding superfamily and FAD-linked reductases C-terminal superfamily (3.4.1 and 4.12.1 superfamilies, respectively). In one group of proteins they appear in enzymes of the oxidoreductase group (e.g. OXDA_CAEEL or PHHY_PSEAE), while in another they are found in a dissociation inhibitor (e.g. GDIA_HUMAN). It should be noted that the proteins are not covered completely by the structural matches, so it is quite possible that the rest of them contain totally different domains that are responsible for the dramatically different functions. However, do these two examples show a rather rare or a more frequent phenomenon? How often do multi-domain proteins, sharing the same structural domain composition, differ in their functions?

In this paper, we attempt to provide a comprehensive answer to this question. This is particularly timely given that most of the unknown proteins in eukaryotic genomes are multi-domain. We use the same approach as in our previous analyses, comparing the sequences of the structural domains in scop to those of SWISS-PROT using BLASTP. We focus on the functional divergence of single and multi-domain proteins, extending previous investigations of single-domain proteins. Also, in comparison to previous work, we focus more on non-enzymatic functions and scop structural superfamilies, instead of folds.

RESULTS

Our Approach to Functional

2000) with $e = 10^{-4}$. We removed the hypothetical and fragment proteins. This resulted in two sets of proteins.

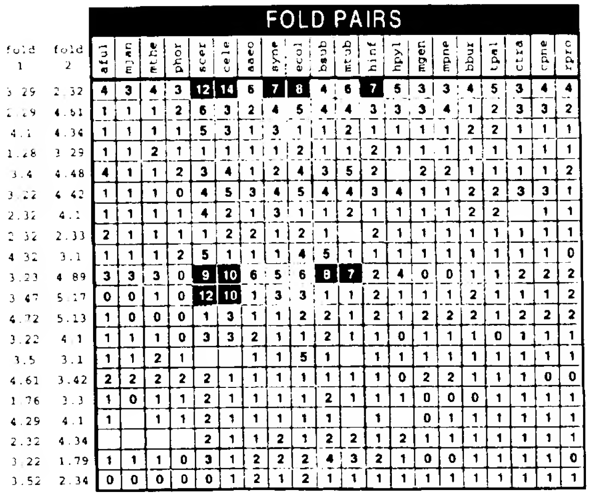
Single-Domain

Of the single-domain matches, only those that were almost completely covered with a match to a single structural domain were selected. (The maximum number of uncovered residues was set at 70 with an additional condition that a maximum of 40 residues on the N-terminal end and 30 residues on the C-terminus were allowed to be uncovered.) These criteria resulted in 1818 single-domain proteins being selected from SWISS-PROT.

Multi-Domain

We selected 4763 multi-domain proteins from SWISS-PROT. All of these matched (in different locations) at least two domains of known structure belonging to different scop superfamilies (see schematic in Figure 1). We also selected a subset of these proteins that have almost their entire length covered by matches with structural domains (allowing again a maximum of 70 uncovered residues). This selection resulted in 2829 proteins being selected from SWISS-PROT. (In all cases, duplicate matches were removed, i.e., a protein at a certain location matches only one structural domain.)

We set out to compare these two sets of proteins for functional divergence. As previously, we divided functions into enzyme and non-enzyme (Hegyi and Gerstein 1999). Enzymatic functions were classified by the EC system (Bairoch 2000). Comparisons of enzymatic functions were treated the same way as in our earlier analyses, that is, if they differ in the first three components of their respective EC numbers, they were considered different. This implied that our analysis dealt with a total of 112 enzymatic functions. Non-enzymatic functions were classified into 508 different categories based on a simple thesaurus we assembled of synonymous keywords drawn from SWISS-PROT description lines. In addition, we created 49 categories for functions that have an enzymatic component but which are not part of the EC system. This gave us a total of 669 functions (112 + 508 + 49). (The list of all the functional categories is described further in Table 2 below, and also can be found on the Web at http://bioinfo.mbb.yale.edu/partslist/func or http://partslist.org/func.)



fold 1	fold 2	aful	mjan	mthe	phor	scer	cele	aaeo	syne	ecol	bsub	mtub	hinf	hpyl	mgen	mpne	bbur	tpal	ctra	cpne	rpro
3.29	2.32	4	3	4	3	12	14	6	7	8	4	6	7	5	3	3	4	5	3	4	4
2.29	4.61	1	1	1	2	6	3	2	4	5	4	4	3	3	3	4	1	2	3	3	2
4.1	4.34	1	1	1	1	5	3	1	3	1	1	2	1	1	1	1	2	2	1	1	1
1.28	3.29	1	1	2	1	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1
3.4	4.48	4	1	1	2	3	4	1	2	4	3	5	2		2	2	1	1	1	1	2
3.22	4.42	1	1	1	0	4	5	3	4	5	4	4	3	4	1	1	2	2	3	3	1
2.32	4.1	1	1	1	1	4	2	1	3	1	1	2	1	1	1	1	2	2		1	1
2.32	2.33	2	1	1	1	1	2	2	1	2	1		2	1	1	1	1	1	1	1	1
4.32	3.1	1	1	1	2	5	1	1	1	4	5	1	1	1	1	1	1	1	1	1	0
3.23	4.89	3	3	3	0	9	10	6	5	6	8	7	2	4	0	0	1	1	2	2	2
3.47	5.17	0	0	1	0	12	10	1	3	3	1	1	2	1	1	1	2	1	1	1	2
4.72	5.13	1	0	0	0	1	3	1	1	2	2	1	2	1	2	2	2	1	2	2	2
3.22	4.1	1	1	1	0	3	3	2	1	1	2	1	1	0	1	1	1	0	1	1	1
3.5	3.1	1	1	2	1			1	1	5	1		1	1	1	1	1	1	1	1	1
4.61	3.42	2	2	2	2	2	1	1	1	1	1	1	1	0	2	2	1	1	1	0	0
1.76	3.3	1	0	1	1	2	1	1	1	1	2	1	1	1	0	0	0	1	1	1	1
4.29	4.1	1		1	1	2	1	1	1	1	1		1		0	1	1	1	1	1	1
2.32	4.34					2	1	1	2	1	2	2	1	2	1	1	1	1	1	1	1
3.22	1.79	1	1	1	0	3	1	2	2	2	4	3	2	1	0	0	1	1	1	1	0
3.52	2.34	0	0	0	0	0	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1

Figure 2 Distribution of multi-domain combinations amongst the genomes. The figure shows the occurrence of multi-domain fold combinations in a number of genomes, indicating its great variability. Each row indicates a particular combination of scop fold pairs (using scop 1.39), where a fold pair is defined as two distinct folds occurring in tandem in a protein. Each column represents a different genome, using the four-letter codes in the PartsList system (Qian et al. 2001): Aaeo, *Aquifex aeolicus*; Aful, *Archaeoglobus fulgidus*; Bbur, *Borrelia burgdorferi*; Bsub, *Bacillus subtilis*; Cele, *Caenorhabditis elegans*; Cpne, *Chlamydia pneumoniae*; Ctra, *Chlamydia trachomatis*; Ecol, *Echerischia coli*; Hinf, *Haemophilus influenzae* Rd; Hpyl, *Helicobacter pylori*; Mthe, *Methanobacterium thermoautotrophicum*; Mjan, *Methanococcus jannaschii*; Mtub, *Mycobacterium tuberculosis*; Mgen, *Mycoplasma genitalium*; Mpne, *Mycoplasma pneumoniae*; Phor, *Pyrococcus horikoshii*; Rpro, *Rickettsia prowazekii*; Scer, *Saccharomyces cerevisiae*; Syne, *Synechocystis* sp.; Tpal, *Treponema pallidum*. The numbers in each intersection cell indicate the number of times the fold pairs occur in a genome. Only the 20 most common fold pair combinations are shown here; the remainder are shown on the Web site (http://partslist.org/func). If a cell is greater than 6, it is shaded black; between 3 and 6, gray; and below 3, white. The blank spaces show instances in which one of the pairs does not occur in the organism at all (indicated by a value of -1 in the data table on the Web site). The fold assignments are done in a fashion consistent with those in PartsList and associated systems (Gerstein 1997; Lin et al. 2000; Drawid et al. 2001; Harrison et al. 2001; Qian et al. 2001).

Overall Distribution of the Matches

Figure 2 shows the most commonly observed multi-domain combinations in a set of recently sequenced genomes. The occurrences of further combinations are available from the Web site. Clearly, the distribution is very skewed, with certain combinations, such as 3.29–2.32, and 2.29–4.61 tending to predominate.

Figure 3 shows the overall distribution of the single-domain and multi-domain matches in the different structural classes. The distribution of matches between enzymes and non-enzymes in multi-domain proteins largely agrees with that in the single-domain proteins. The multi-domain matches follow the overall tendency of the alpha/beta folds to be associated with enzymes to a larger extent and the all-alpha and small folds with non-enzymes. However, the values for the multi-domain matches are generally less extreme than tural classes compared to the single-domain matches. Altogether, there are more enzymes than non-enzymes among the multi-domain proteins (2805 enzymes vs. 1958 non-enzymes) whereas for single-domain proteins, the opposite is true (850 enzymes vs. 968 non-enzymes).

Table 1 summarizes the distribution of superfamilies and superfamily combinations among the major functional classes, i.e. whether they have only enzymatic, only non-enzymatic or both enzymatic and non-enzymatic functionality. Altogether, 215 superfamilies were found in single-domain proteins and 310 in multi-domain ones. As 70 superfamilies were found in both, altogether 455 distinct structural superfamilies matched a SWISS-PROT protein with our required coverage criteria (described above). Similarly, we apportioned the 281 superfamily combinations observed in multi-domain

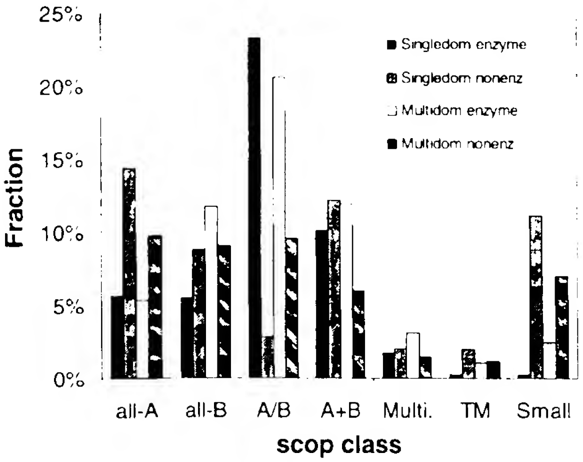


Figure 3 Distribution of proteins amongst broad structural and functional classes; the distribution of the matches among the seven structural and two functional classes in single- and multi-domain proteins. The single-domain and multi-domain matches each total 100%, independently of each other. The horizontal axis indicates the seven scop classes, which are (from 1 to 7): all-alpha, all-beta, alpha/beta, alpha + beta, multi-domain, membrane, and small protein.

to almost threefold (135 vs. 56). This agrees with the notion that most enzymes are multi-domain. Another difference between single and multi-domain proteins appears in the ratio of superfamilies with a single function compared to multifunctional ones. As it is apparent from Table 1, about a quarter of the superfamilies matched single-domain proteins with different functions (55 of 215), whereas in the multi-domain proteins, this ratio increased to more than a third (119 of 310).

Single-Domain Proteins

Table 2 lists the two functionally most diverse structural superfamilies in single-domain proteins with some representative functions. The most diverse superfamily, the 3.38.1 Thioredoxin-like, has 11 different functions associated with it, most of them with an oxidoreductase mechanism. For instance, THIO_BPT4 is a small disulphide-containing thioredoxin that serves as a general disulphide oxidoreductase, while TDX2_BRUMA is almost twice as long (199 aa) and serves as a thiol-specific antioxidant that acts against sulfur-containing radicals. Another interesting example of functional diversity is provided by the Scorpion toxin-like superfamily (7.3.6). While BRAZ_PENBA is a small protein that is known to be 2000 times sweeter than sucrose, the other members of the superfamily are associated with different host-defense mechanisms. In insects the superfamily possesses antifungal activity (DMYC_DROME) or acts as a toxin (SCX5_BUTEU). Interestingly, in plants it can also act as an antifungal (AF2B_SINAL) or as an inhibitor of insect alpha-amylases (SIA1_SORBI). It appears that many single-domain proteins are toxins or allergens, or are related in other ways to a host-defense response.

Based on the data we can also determine the probability of two single-domain proteins that match domains in the same superfamily category also carrying out the same function. Using Bayes' theorem:

$$P(F|S) = P(F)P(S|F)/((P(F)P(S|F) + P(\bar{}F)P(S|\bar{}F)) \qquad (1)$$

where S is the probability that two proteins share the same superfamily, F is the probability that two proteins have the same function, and $\bar{}F$ is the probability that two proteins do not have the same function. Rearranging and simplifying the equation we get:

$$P(F|S) = 1/(1 + N(S,\bar{}F)/(N(S,F)) \qquad (2)$$

where N is the number of times that the two events in the parentheses occur together in our database of 1818 single-domain proteins. This results in

$$P(F|S) = 1/(1 + 8501/12516) = 68\%.$$

That is, the probability that two single-domain proteins that have the same superfamily structure have the same function (whether enzymatic or not) is about 2/3.

Multi-Domain Proteins

Table 3 lists the combinations of superfamilies that have been associated with the greatest number of different functions in multi-domain proteins, with representative entries in SWISS-PROT. The combination with the greatest number of different functions is that of 1.95.1 and 7.33.1. Although it has twice as many different functions as the most diverse superfamily in

Table 1. **Functional Distribution of Single-domain, Multi-domain Superfamilies, and Multi-domain Combinations**

	Single-domain superfamilies		Multi-domain superfamilies		Multi-domain sfam combinations	
	Single function	Multiple function	Single function	Multiple function	Single function	Multiple function
Enzymatic	82	11	135	42	151	16
Nonenzymatic	78	23	56	30	70	27
Both functions	—	15	—	47	—	17
Total	160	55	191	119	221	60

The basic functional distribution of the superfamilies in single- and multi-domain proteins and the functional distribution of multi-domain combinations are shown. The first row lists the number of superfamilies that were associated only with [illegible]

Table 2. Most Versatile Single-Domain Superfamilies

No. func	No. prot	Sfam comb	Function	SWISS-PROT ID	SWISS-PROT function
11	69	3.38.1	E1.11.1	GSHP_RAT	Plasma Glutathione Peroxidase (1.11.1.9)
			263#	DYL5_CHLRE	Dynein, Flagellar Outer Arm–*C. reinhardtii*
			D260#	BSAA_BACSU	Glutathione Peroxidase Homolog Bsaa
			268#	REHY_TORRU	Rehydrin–*Tortula ruralis* (Moss)
			266#	PHOS_HUMAN	Phosducin (33 Kd Phototransducing Protein)
			269#	REHY_ORYSA	Rad24 Protein–*Oryza sativa* (Rice)
			272#	THIO_BPT4	Thioredoxin (Bacteriophage T4)
			D271#272#	TDX2_BRUMA	Thioredoxin Peroxidase 2
			261#	BTUE_ECOLI	Vitamin B12 Transport Periplasmic Protein Btue
10	28	7.3.6	342#	BRAZ_PENBA	Brazzein–*Pentadiplandra brazzeana*
			376#336#	SCKK_TITSE	Neurotoxin Ts-Kapa (Tsk)–(Brazilian scorpion)
			341#356#	AF2B_SINAL	Cysteine-Rich Antifungal Protein 2b (Afp2b)
			343#	DEFA_ZOPAT	Defensin, Isoforms B And C–*Zophobas atratus*
			361#	DMYC_DROME	Drosomycin Precursor (Cysteine-Rich Peptide)
			361#376#	SCX5_BUTEU	Insectotoxin I5a–(Lesser Asian scorpion)
			336#	SCX3_LEIQH	Leiuropeptide Iii–(Scorpion)
			203#	SIA1_SORBI	Small-Pr Inhibitor Of Insect Alpha-Amylases
7	34	4.79.3	310#	AB18_PEA	Aba-Responsive Protein Abr18–Garden Pea
			311#	DRR3_PEA	Disease Resistance Response Protein Pi49
			231#	MPAA_CORAV	Major Pollen Allergen Cor A 1,–Eu. Hazel
			312#	L18B_LUPLU	Protein L1r18b (Llpr10.1b)
			E3.1.-	RNS2_PANGI	Ribonuclease 2 (3.1.–/–)–Panax Ginseng
			314#	SAM2_SOYBN	Stress-Induced Protein Sam22
7	43	1.26.1	184#	CSF2_SHEEP	Colony-Stimulating Factor
			381#564#184#	IL4_RAT	Interleukin-4 (B-Cell Igg Diff. Factor)
			185#	LIF_HUMAN	Leukemia Inhibitory Factor (Lif)
			187#	PRL_ANGAN	Prolactin Precursor (Prl)–
			186#	PLF3_MOUSE	Proliferin 3 Mitogen-Regulated
			188#	SOMA_PAROL	Somatotropin (Growth Hormone)

The most versatile superfamilies in single-domain proteins as determined from their functional description in SWISS-PROT, with some representatives. The keyword combinations in the fourth column were based either on the first three components of their EC numbers (for enzymes) or derived automatically by comparing the DE description line of SWISS-PROT entries to a list of synonymous keywords at http://bioinfo.mbb.yale.edu/partslist/func. A keyword number starting with a D indicates an enzyme that does not have an assigned EC number in its description in SWISS-PROT.

the single-domain proteins (22 vs. 11, respectively), careful examination reveals that all the proteins in this category are DNA-binding and most of them act as hormone receptors.

The second entry listed in the table is the combination of the 3.4.1 and 4.48.1 superfamilies associated with the FAD/NAD(P)-linked reductases. It is an all-enzymatic combination and always carries out an oxido-reductase function. All the proteins in this category are completely covered by matches with these two superfamilies. The 1.78.1–2.1.1 hemocyanin-immunoglobulin combination seems also to be fairly conserved; although the proteins in this category are called by eight different names, most of them turn out to be extracellular larval storage proteins, except for the copper-containing oxygen carrier hemocyanin itself (HCY_PALVU).

Following the same logic, we can also determine the probability that two proteins that have the same superfamily combination share the same function, viz:

$$P(F|S) = 1/(1 + 32242/134230) = 81\%$$

almost complete coverage with exactly the same type and number of superfamilies, following each other in the same order. The probability that the functions are the same in this case was 91%, a considerably higher value than above. However, if two multi-domain proteins share only a single superfamily, the probability that they share the same function drops to only 35%! This greater functional certainty from sharing a combination of superfamilies rather than just one is also reflected in Table 1. While one-fourth of the single-domain proteins and one-third of singularly matching superfamilies in multi-domain proteins have multiple functions, only about one-fifth of the multi-domain *combinations* possess multiple functions (60 of 281). It is also clear from the data that domains in larger proteins often lose their original function and no longer have an autonomous function.

Seventy Common Superfamilies and Their Functions Compared in Single-Domain and Multi-Domain Proteins

Table 3. Most Versatile Superfamily Combinations in Multi-Domain Proteins

No. func	No. prot	Sfam comb.	Function	SWISS-PROT ID	SWISS-PROT function
22	176	1.95.1/7.33.1	29#	THB_RANCA	Thyroid Hormone Receptor Beta
			10#	HNF4_DROME	Transcription Factor HNF-4 Homolog
			31#32#	EAR2_MOUSE	V-Erba Related Protein Ear-2
			29#30#	ECR_MANSE	Ecdysone Receptor (Ecdysteroid Receptor)
			32#	ERBA_AVIER	Erba Oncogene Protein
			556#564#35#	NGFI_XENLA	Nerve Growth Factor Induced Protein I-B
			576#	NR42_HUMAN	Immediate-Early Response Protein Not
			36#	PPAT_HUMAN	Peroxisome Proliferator Activated Receptor
			37#	RXTG_CHICK	Retinoic Acid Receptor RXR-Gamma
			38#	TLL_DROVI	Tailless Protein
8	54	3.4.1/4.48.1	E1.8.2	DHSU_CHRVI	Sulfide Dehydrogenase (1.8.2.–)
			E1.8.1	DLDH_ZYMMO	Dihydrolipoamide Dehydrogenase (1.8.1.4)
			E1.6.4	TYTR_TRYCR	Trypanothione Reductase (1.6.4.8) (Tr)
			E1.16.1	MERA_STRLI	Mercuric Reductase (1.16.1.1)
			E1.6.99	NAOX_MYCPN	Probable NADH Oxidase (1.6.99.3) (Noxase)
8	23	1.78.1/2.1.1	19#	ARYB_MANSE	Arylphorin Beta Subunit–(Tobacco Hornworm)
			20#	CRPI_PERAM	Allergen Cr-Pi Precursor–(American Cockroach)
			21#427#	HCY_PALVU	Hemocyanin–(European Spiny Lobster)
			22#	HEXA_BLADI	Hexamerin Precursor–(Tropical Cockroach)
			23#	JSP1_TRINI	Acidic Juvenile Hormonne–Suppressible Protein
			24#	LSP2_DROME	Larval Serum Protein 2 Precursor (LSP-2)
			546#25#	SSP1_BOMMO	Sex-Specific Storage–Protein 1

Note that the combination with the greatest number of different functions is that of 1.95.1 and 7.33.1. Careful examination reveals that all the proteins with this combination are DNA-binding and most of them act as various hormone receptors. In particular, HNF4_DROME and NR42_HUMAN also have transcription activator functions. Note that these two proteins are considerably longer than the others in this group and are not covered completely by structural matches: A large C-terminal and a large N-terminal portion are left uncovered, respectively.

multi-domain proteins. These are listed in Table 4; 12 of them have enzymatic function, supporting the notion that enzymes are more conserved during evolution than non-enzymes. The two non-enzymatic superfamilies are the 4.29.1 ribosomal superfamily and the 5.4.1 superfamily in penicillin-binding proteins.

Table 5 presents several examples of the converse situation, shared superfamilies that have different functions in single and multi-domain proteins. Comparing parts A and B of the table highlights the fact that although both superfamilies in a multi-domain protein are often present in single-domain form as well, the functions in the different settings are only vaguely related. One example is the combination of the lipocalin superfamily (2.45.1) with that of the BPTI-like or Kunitz inhibitor (7.7.1), which in higher organisms forms a complex protein called alpha-1-microglobulin (AMBP_RAT). Another interesting example is the combination of the 2.5.1 Cupredoxin (occurring in the single-domain blue-copper protein, SOXE_SULAC) and the 6.5.1 Membrane all-alpha (single-domain representative: BACT_HALVA, a sensory rho-

Table 4. Superfamilies With the Same Function in Single- and Multi-Domain Proteins as Determined from Their Keyword Combination or First Three Components of Their EC Numbers

		Single-domain proteins		Multi-domain proteins	
Sfam	Function	SWISS-PROT ID	SWISS-PROT function	SWISS-PROT ID	SWISS-PROT function
1.81.1	E3.2.1	GUNY_ERWCH	Endoglucanase (3.2.1.4)	AMYG_NEUCR	Glucoamylase Precursor (3.2.1.3)
2.66.2	E3.5.1	URE2_YERPS	Urease Beta (3.5.1.5)	URE1_HELPY	Urease Alpha Subunit (3.5.1.5)
3.17.2	E6.3.5	NADE_MYCPN	NAD(+) Synthetase (6.3.5.1)	GUAA_YEAST	GMP Synthase (6.3.5.2)
3.37.1	E3.1.3	PTP2_NPVOP	Protein-Tyrosine Phosphatase 2 (3.1.3.48)	PTNB_RAT	Protein-Tyrosine Phosphatase (3.1.3.48)
3.67.1	E4.2.1	TRPB_VIBPA	Tryptophan Synthase (4.2.1.20)	TRP_YEAST	Tryptophan Synthase (4.2.1.20)
4.19.1	E5.2.1	FKB1_METJA	Peptidylprolyl *Cis-Trans* Isomerase (5.2.1.8)	FKB7_WHEAT	70 Kd Peptidylprolyl Isomerase (5.2.1.8)
4.2.1	E3.2.1	LYCV_BPP2	Lysozyme (3.2.1.17)	CHIX_PEA	Endochitinase Precursor (3.2.1.14)
4.29.1	85#	RS5 ACYKS	30s Ribosomal Protein S5	RS5 TREPA	30s Ribosomal Protein S5
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Table 5. Examples of Superfamilies Present in Both Single- and Multi-Domain Proteins, Carrying out Different Functions

Table 5A. Single-Domain Proteins

Sfam	Funct #	SWISS-PROT ID	SWISS-PROT function
1.25.1	352# 183# E1.17.4 192#	FTN2_HAEIN NIGY_DESVH RIR4_YEAST NLP_HAEIN	Ferritin-like Protein 2 Nigerythrin (Ribonucleotide Reductase) (1.17.4.1) Ner-like Protein Homolog
1.4.3	196#	H1A_PLADU	Histone H1A, Sperm
1.81.2	E2.5.1	PFTB_PEA	Farnesyltransferase Beta Su (2.5.1.–)
2.45.1	226# 227# 228#412# 229# E5.3.99 230#421#	ERBP_RAT FAB3_CAEEL NGAL_MOUSE NP4_RHOPR PGHD_HUMAN VNS1_MOUSE	Epididymal-Tetinoic Acid Binding Protein Fatty Acid-Binding Protein Homolog 3 Neutrophil Gelatinase-Assoc. Lipocalin Nitrophorin 4 Precursor Prostaglandin-H2 D-Isomerase (5.3.99.2) Vesomeral Secretory Protein I
2.5.1	231# 232#427#	MPA3_AMBEL SOXE_SULAC	Pollen Allergen AMB A 3 (AMB A Iii) Sulfocyanin (Blue Copper Protein)
3.14.2	373#	RRF1_DESVH	Rrf1 Protein
3.29.1	E6.3.4 E2.7.4 D259# E2.7.1	PURA_CAEEL KTHY_YEAST VA57_VACCV KITH_VZVW	Adenylosuccinate Synthetase (6.3.4.4) Thymidylate Kinase (2.7.4.9) Guanylate Kinase Homolog Thymidine Kinase (2.7.1.21)
3.47.1	275# 276#	MBL_BACSU MREB_BACSU	MBL Protein Rod Shape-determining Protein Mreb
3.48.1	E3.1.3	PPA5_YEAST	Repressible Acid Phosphatase (3.1.3.2)
3.81.1	D281# 282#	AMIC_PSEAE LUXP_VIBHA	Aliphatic Amidase Expression-Regulator LUXP Protein Precursor
4.103.1	E2/4/2	TOX1_BORPE	Pertussis Toxin Su 1 (2.4.2.–)
4.105.1	291#	LECC_POLMI	Lectin–Polyandrocarpa Misakiensis
4.11.5	295#	TERP_PSESP	Terpredoxin
4.19.1	E5.2.1	FKB1_METJA	Pept-Prolyl *Cis-Trans* Isomerase (5.2.1.8)
6.5.1	E3.6.1 540#325#	ATPL_VIBAL BACT_HALVA	ATP Synthase (3.6.1.34) (Lipid-binding) Sensory Rhodopsin II (Sr-Ii)
7.35.4	E1.9.3 345#	COXB_RAT DESR_DESBI	Cytochrome C Oxidase (1.9.3.1) (Via*) Desulforedoxin (Dx)
7.7.1	349#	TAP_ORNMO	Tick Anticoagulant Peptide

(*Table continues on following page.*)

dopsin) superfamilies into a component of the respiratory chain, cytochrome C oxidase II (COOX_ZOOAN). All these examples demonstrate the evolutionary advantage of a domain fusion event, which creates a function that is more complex than either of the components.

Gerstein 1999; Wilson et al. 2000). Figure 4 shows a similar graph with the calculations extended to multi-domain proteins. The figure shows that the functional divergence of a single domain in multi-domain proteins dramatically increases, more than twofold, compared to the single-domain

Table 5B. Multi-Domain Proteins

Sfam Comb.	Funct#	SWISS-PROT ID	SWISS-PROT function
1.25.1/7.35.4	104#	RUBY_METJA	Putative Rubrerythrin
1.32.1/3.81.1	11# 12# 581#11# 582#11#	PURR_HAEIN DEGA_BACSU SCRR_STRMU REGA_CLOAB	Purine Nucleotide Synthesis Repressor Degradation Activator Sucrose Operon Repressor Transcription Regulatory Protein Rega
1.4.3/3.14.2	10# 11# 13# 190# 366#	SKN7_YEAST VIRG_AGRT5 RGX3_MYCTU PFER_PSEAE PETR_RHOCA	Transcription Factor Skn7 (Pos9 Protein) Virg Regulatory Protein Sensory Transduction Protein REGX3 Transcriptional Activator Protein Pfer Petr Protein
2.45.1/7.7.1	203#153#	HC_RAT	Alpha-1-Microglobulin/Trypsin Inhibitor
2.5.1/6.5.1	E1.9.3	COX2_ZOOAN	Cytochrome C Oxidase Ii (1.9.3.1)
3.29.1/3.48.1	E2.7.1	F26_RANCA	6-Phosphofructo-2-Kinase (2.7.1.105)
3.47.1/5.17.1	1# 1#83#	YED0_YEAST GR73_MAIZE	Heat Shock Protein 70 Homolog YEL030w Ig-Binding Protein

DISCUSSION

Here we built on our previous studies on the relationship between protein structure and function to develop new results related to multi-domain proteins. Throughout the paper, we focused on superfamilies instead of folds, as the members of a superfamily are presumably of common evolutionary origin (Murzin et al. 1995).

We found that the 4763 multi-domain and 1818 single-domain proteins that met our selection criteria have about the same distribution of structural classes, with more enzymatic functions associated with the alpha/beta structural classes and more non-enzymatic ones with the all-alpha and small classes. We identified more than three times as many multi-domain proteins that were enzymes than single-domain ones (2805 and 850, respectively) and, conversely, about twice as many multi-domain proteins as single-domain ones that were non-enzymes (1958 vs. 968).

We focused on the functional divergence of the two groups and found that about a quarter of the superfamilies in single-domain proteins are associated with multiple functions, whereas only about a fifth of the multi-domain superfamily combinations are. Therefore, we can conclude that a combination of specific superfamilies results in a more specific functional assignment for a particular protein. However, about one-third of the superfamilies in the multi-domain proteins were associated with multiple functions, underlining the lesser autonomy of a domain function in multi-domain protein.

This latter finding was also supported by the difference in functional divergences between the two groups of proteins based on particular sequence similarities between the domains and SWISS-PROT proteins. As is shown in Figure 4, the average functional divergence of a single domain is much larger (more than twofold) in multi-domain proteins than in [illegible] was rather surprising to us, and should be taken into consideration in functional characterization and annotation of new gene products. When the functions were related in single- and multi-domain proteins, we could observe an increasing functional complexity with the appearance of large multi-domain proteins.

Altogether, with the recent sequencing of the human genome and the genomes of other model organisms, we hope

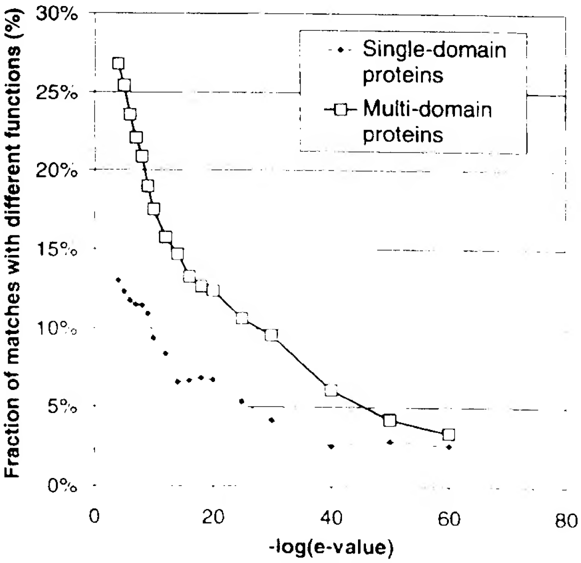


Figure 4 Divergence in function with respect to sequence [illegible]

that this work can contribute to the successful annotation of the individual gene products, and will help to avoid some pitfalls associated with the functional characterization of large, complex proteins.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389-3402.

Bairoch, A. 2000. The ENZYME database in 2000. *Nucleic Acids Res.* **28:** 304-5.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45-8.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5:** 823-826.

Devos, D. and Valencia, A. 2000. Practical limits of function prediction. *Proteins* **41:** 98-107.

Drawid, A. and Gerstein, M. 2000. A Bayesian system integrating expression data with sequence patterns for localizing proteins: Comprehensive application to the yeast genome. *J. Mol. Biol.* **301:** 1059-1075.

Eisenstein, E., Gilliland, G. L., Herzberg, O., Moult, J., Orban, J., Poljak, R. J., Banerjei, L., Richardson, D. and Howard, A. J. 2000. Biological function made crystal clear - annotation of hypothetical proteins via structural genomics. *Curr. Opin. Biotechnol.* **11:** 25-30.

Gelbart, W. M., Crosby, M., Matthews, B., Rindone, W. P., Chillemi, J., Russo Twombly, S., Emmert, D., Ashburner, M., Drysdale, R. A., et al. 1997. FlyBase: A *Drosophila* database. The FlyBase consortium. *Nucleic Acids Res.* **25:** 63-6.

Gerstein, M. 1997. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274:** 562-76.

———. 1998. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des.* **3:** 497-512.

Harrison, P. Echols, N. and Gerstein, M. 2001. Digging for dead genes: An analysis of the characteristics of the pseudogene population in the *C. elegans* genome. *Nucleic Acids Res.* **29:** 818-830.

Hegyi, H. and Gerstein, M. 1999. The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288:** 147-164.

Lin, J. and Gerstein, M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels. *Genome Res.* **10:** 808-818.

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C. and Thornton, J. M. 1998. Protein folds and functions. *Structure* **6:** 875-884.

Murzin, A., Brenner, S. E., Hubbard, T. and Chothia, C. 1995. SCOP: A structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536-540.

Orengo, C. A., Pearl, F. M., Bray, J. E., Todd, A. E., Martin, A. C., Lo Conte, L. and Thornton, J. M. 1999. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.* **27:** 275-279.

Pawlowski, K., Jaroszewski, L., Rychlewski, L. and Godzik, A. 2000. Sensitive sequence comparison as protein function predictor. *Pac. Symp. Biocomput.* 42-53.

Pearson, W. R. 1994. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.* **25:** 365-389.

Qian, J., Stenger, B., Wilson, C., Lin, J., Jansen, R., Krebs, W., Alexandrov, V., Echols, N., Teichmann, S., Park, J. et al.. 2001. PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res.* **29:** 1750-1764.

Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A. and Sternberg, M. J. 1997. Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *J. Mol. Biol.* **269:** 423-439.

Shah, I. and Hunter, L. 1997. Predicting enzyme function from sequence: A systematic appraisal. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 276-283.

Shapiro, L. and Harris, T. 2000. Finding function through structural genomics. *Curr. Opin. Biotechnol.* **11:** 31-5.

Stawiski, E.W., Baucom A.E., Lohr S.C., and Gregoret L.M. 2000. Predicting protein function from structure: Unique structural features of proteases. *Proc. Natl. Acad. Sci.* **97:** 3954-8.

Thornton, J. M., Orengo, C. A., Todd, A. E. and Pearl, F. M. 1999. Protein folds, functions and evolution. *J. Mol. Biol.* **293:** 333-342.

Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. and Orengo, C. A. 2000. From structure to function: Approaches and limitations. *Nat. Struct. Biol.* **7 Suppl:** 991-994.

Todd A.E., Orengo C.A., and Thornton J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307:** 1113-1143.

Wilson, C. A., Kreychman, J. and Gerstein, M. 2000. Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297:** 233-249.

Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J. S., Skolnick, J. and Godzik, A. 1999. From fold predictions to function predictions: Automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* **8:** 1104-1115.

Received February 7, 2001; accepted in revised form June 19, 2001.